Advances in Artificial Intelligence

Artificial Intelligence Applications in Biomedicine

Guest Editors: Panayiotis Vlamos, Konstantinos Lefkimmiatis, Catalina Cocianu, Luminita State, and Zhiyuan Luo

Artificial Intelligence Applications in Biomedicine

Artificial Intelligence Applications in Biomedicine

Guest Editors: Panayiotis Vlamos, Konstantinos Lefkimmiatis, Catalina Cocianu, Luminita State, and Zhiyuan Luo

Editorial Board

Contents

Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2013, Article ID 219137, 2 pages
http://dx.doi.org/10.1155/2013/219137

Editorial

Artificial Intelligence Applications in Biomedicine

Panayiotis Vlamos,[1] Konstantinos Lefkimmiatis,[2] Catalina Cocianu,[3] Luminita State,[4] and Zhiyuan Luo[5]

[1] *Department of Informatics, Ionian University, Palaia Anaktora, 49100 Corfu, Greece*
[2] *Brigham and Women's Hospital and the VA Boston Healthcare System, Harvard Medical School, West Roxbury VAMC, 1400 VFW Parkway Room 2B111, West Roxbury, MA 02132, USA*
[3] *Department of Informatics in Economy, Faculty of Economic Cybernetics, Statistics and Informatics, Bucharest Academy of Economic Studies, 15-17 Calea Dorobantilor, District 1, 71131 Bucharest, Romania*
[4] *Department of Mathematics and Computer Science, University of Pitesti, Street Targu din Vale, No. 1, Arges, 110040 Pitesti, Romania*
[5] *Department of Computer Science, Royal Holloway University of London, Egham, Surrey TW20 0EX, UK*

Correspondence should be addressed to Panayiotis Vlamos; vlamos@ionio.gr

Received 22 November 2012; Accepted 22 November 2012

"Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it."—The Dartmouth Summer Research Project on Artificial Intelligence.

Recent technological advances in medical informatics and biomedicine facilitated the development of complex biomedical systems including innovated clinical and computer-based decision support systems, knowledge acquisition and management, medical imaging, computational intelligence in bioclinical medicine, molecular medicine, and healthcare organizational aspects. Artificial intelligence (AI) has a great impact on the fields of biology, biotechnology, and medicine in general and can be implemented in real world applications through machine learning techniques, neural computing, expert systems, fuzzy logic, genetic algorithms or Bayesian modelling. This issue of Artificial Intelligence Applications in Biomedicine (AIAB) compiles five innovating research articles, concerning machine learning methodologies, AI support systems in patient monitoring, genetic algorithms as well as optimization techniques and semantics.

The problem of classification, prediction, and diagnostic (CPD) errors is addressed in E. Eiland et al. paper, as well as the importance of relative class size (rCS) to end user problems. Authors made the first step in identifying the structure of CPD problems faced by end users and described the way that CPD evaluation measures and tools are relevant to end users. They have also identified measures that are efficacious for end users and shown how joint probability table (JPT) normalization and JPT tuning are useful for end user CPD evaluation.

A. Alexiou and P. Vlamos designed a new genetic algorithm (GA) for the representation of mitochondrial population through the framework of cultural algorithms (CA). It is well known that mitochondrial dysfunctions are highly associated with neurodegenerative diseases and related disorders; therefore, authors provide a combined simulation procedure for the optimization problem of ATP production in cells. The proposed algorithm is related to the main mitochondrial dynamics of complete fusion, transient fusion, fission and motility and the way that these operations can affect healthy mitochondrial population as an early symptom of neurodegenerative diseases.

An integrated system based on web applications, smartphones and an interconnection to Microsoft HealthVault platform is developed from E. Sourla et al. for monitoring chronic cardiology patients and early notifying and optimizing the process management of an emergency cardiologic incident. The decision system (DSS) supports cardiologic patient modules and is based on common cardiology diseases

and fuzzy logic. The proposed AI system, CardioSmart365 is a complex expert system that combines DSS and web services, concerning patients, medical doctors, everyday clinical practice, research and science, and Healthcare Systems.

Additionally, the basin hopping (BH) framework has been explored in A. Shehu et al. research paper, as a powerful global optimization tool for systems with multiple variables and modalities like biological macromolecules (proteins). Authors took into consideration different applications of importance in computational structural biology in order to demonstrate the importance of BH framework and specific characteristics such as adjacency of local minima and its relation to the quality of the reported global minimum.

Finally Z. Majkic extended both Montague's and Bealer's approaches for the intentional first-order logic (FOL) and design also a new extensional algebra for the FOL and the commutative homomorphic diagram that express the generalization of the Tarskian theory of truth for the FOL into the Frege/Russell's theory of meaning.

In conclusion, we hope these papers to enrich our readers and researchers in both theoretical and applied modern scientific fields, with respect to the strong relationship between artificial intelligence, information technology and mathematical modelling in biology and medicine.

Panayiotis Vlamos
Konstantinos Lefkimmiatis
Catalina Cocianu
Luminita State
Zhiyuan Luo

Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2013, Article ID 427958, 22 pages
http://dx.doi.org/10.1155/2013/427958

Research Article

Efficacious End User Measures—Part 1: Relative Class Size and End User Problem Domains

E. Earl Eiland and Lorie M. Liebrock

Computer Science and Engineering Department, New Mexico Institute of Mining and Technology, 801 Leroy Place, Socorro, NM 87801, USA

Correspondence should be addressed to E. Earl Eiland; eee@nmt.edu

Received 29 June 2012; Accepted 28 October 2012

Academic Editor: Konstantinos Lefkimmiatis

Biological and medical endeavors are beginning to realize the benefits of artificial intelligence and machine learning. However, classification, prediction, and diagnostic (CPD) errors can cause significant losses, even loss of life. Hence, end users are best served when they have performance information relevant to their needs, this paper's focus. Relative class size (rCS) is commonly recognized as a confounding factor in CPD evaluation. Unfortunately, rCS-invariant measures are not easily mapped to end user conditions. We determine a cause of rCS invariance, joint probability table (JPT) normalization. JPT normalization means that more end user efficacious measures can be used without sacrificing invariance. An important revelation is that without data normalization, the Matthews correlation coefficient (MCC) and information coefficient (IC) are not relative class size invariants; this is a potential source of confusion, as we found not all reports using MCC or IC normalize their data. We derive MCC rCS-invariant expression. JPT normalization can be extended to allow JPT rCS to be set to any desired value (JPT tuning). This makes sensitivity analysis feasible, a benefit to both applied researchers and practitioners (end users). We apply our findings to two published CPD studies to illustrate how end users benefit.

1. Introduction

Biological compounds and systems can be complex, making them difficult to analyze and challenging to understand. This has slowed applying biological and medical advances in the field. Recently, artificial intelligence and machine learning, being particularly effective classification, prediction and diagnostic (CPD) tools, have sped applied research and product development. CPD can be described as the act of comparing observations to models, then deciding whether or not the observations fit the model. Based on some predetermined criterion or criteria, a decision is made regarding class membership ($x \in A$ or $x \notin A$). In many domains, class affiliation is not the end result, rather it is used to determine subsequent activities. Examples include medical diagnoses, bioinformatics, intrusion detection, information retrieval, and patent classification. The list is virtually endless. Incorrect CPD output can lead to frustration, financial loss, and even death; correct CPD output is important. Hence, a number of CPD algorithms have been developed and the field continues to be active.

Characterizing CPD effectiveness, then, is necessary. For example, CPD tool developers need to know how their particular modification affects CPD performance, and practitioners want to make informed choices between CPD options before deploying a tool in the field.[1] Jamain and Hand, summarizing their results in a classifier meta-analysis, comment:

> *The real question a user generally wants to answer is "which classification methods [are] best for me to use on my problem with my data . . ." [1].*

This question has not been addressed in studies we have read. Indeed, Jamain and Hand generalize the sentiment of R.P.W Duin's comment regarding comparing automated, heavily parametrized classifiers.

> *It is difficult to compare these types of classifiers in a fair and objective way [2].*

Seemingly, the research community has viewed the end user's need as too complex to address. Thus, for the most part, researchers have focused on addressing their own needs. End user issues, when discussed, have been constrained to specific problem domains. It might be fair to state that each end user's need is, in some way, unique. However, that does not mean that the apparent complexities faced by end users cannot be identified and managed. Ideally, a means of satisfying end user needs without sacrificing researcher needs will emerge. At a minimum, it should be possible for end users to be enlightened regarding measure suitability (which measures best quantify how a CPD will impact their situation). This paper is a first step in identifying a general structure of CPD problems[2] faced by end users and using that structure to identify CPD evaluation measures and tools relevant to end users. To the extent that research studies present CPD performance information by which end users can estimate impact in their situation, the studies provide improved service to the end user.

Our primary focus is on summary statistics. In the current context, summary statistics are formulae that take measurement suite elements as input[3] and output a single value which represents the target CPD's overall quality. However, because multiple values are condensed into a single value, information is lost. To the extent essential information is retained, the summary statistic can prove useful for CPD evaluation. A key characteristic of summary statistics is that they are not monotonic; they have optima. Useful summary statistic optima indicate overall classifier quality. Ideally, these summary statistics also quantify some aspect of classifier output efficacious to end users. End users can directly use such values to estimate how the CPD will impact their situation.

As presented by Hand [3], measurement theory distinguishes between two entity attribute types: *intrinsic*, those that are part of an entity's definition (e.g., density or mass) and *extrinsic*, those that are expressions of the entity's interaction with the environment (e.g., weight). Attributes such as density and weight can be quantified, so we can also talk about intrinsic and extrinsic measures. When reported in joint probability tables (JPTs), CPD output is partitioned into four distinct categories: T_+, F_+, F_-, and T_-. After any dataset has been tested, the final object count in each category is influenced by the environmental factors rCS and boundary (B). (rCS is the relative sizes of the classes in the test set (rCS = $\overline{Y}/Y$). B is an n element vector that defines a "surface" that encloses one class, for example, "class A." In every case, there will also be an optimum boundary[4] (B^*). All elements outside that surface are in class "$\overline{A}$", rather than class "A." Because T_+, F_+, F_-, and T_- are sensitive to rCS and B, they are extrinsic measures.

1.1. Nomenclature. Although this paper applies well-established stochastic concepts, not all discussions use the same terminology. To avoid confusion, we define our lexicon for quantities measured (each being the size of the defined set):

T_+: correctly identified events in class A, the "class of interest" (if such a class exists);

T_-: correctly identified events of class $\overline{A}$, the other class;

F_+: class $\overline{A}$ events incorrectly flagged as class A;

F_-: class A events incorrectly flagged as class $\overline{A}$;

Z: events flagged as class A;

$\overline{Z}$: events flagged as class $\overline{A}$;

Y: actual class A events in the data set;

$\overline{Y}$: actual class $\overline{A}$ events in the data set;

N: the data set;

Table 1: Values in the lexicon are often organized into a joint probability table (JPT), such as this.

		Actual target classification		
		A	$\overline{A}$	Totals ↓
Test result	Positive	T_+	F_+	Z
	Negative	F_-	T_-	$\overline{Z}$
	Totals	Y	$\overline{Y}$	N

These values are often presented in an JPT as in Table 1. When appropriate, these symbols will also be used to represent populations. Context will determine whether a quantity or a population is being referenced.

End users are interested in how a process will function in their environment, so they need measures sensitive to extrinsic factors. From a purely academic perspective, the goal for many researchers is to characterize the CPD process independent of extrinsic factors; thus, they want intrinsic measures. Presumably because of the immediacy of the need, significant progress has been made in identifying and characterizing intrinsic measures.[5] We are interested in extrinsic measures useful for end users; little attention has been paid to their needs.

Because of the disparity between researcher and end user needs, providing for end user needs requires careful consideration. A researcher is interested solely in CPD performance; effects caused by external factors must be accounted for, if not eliminated. In contrast, end users need to incorporate external factors, not compensate for or eliminate them. Thus, in order to have research reports that are readily applicable by end users may require providing values that hold little relevance for researchers. We propose an "end user efficacious" measure suite and a means by which end users can tailor research results to their specific environment.

This study builds on Sokolova et al. and other CPD summary statistic characterization studies [4–14]. A challenge categorical problem evaluators face, when comparing to CPD results reported by others, is adjusting for data set effects. One of the major data set issues is that test sets used may well have different rCSs, with different applicability and/or utility. This can cloud results. As an example, we ran a CPD on two test sets drawn from the same class populations. Since both the class source populations and CPD were the same for each test, one would expect statistically indistinguishable output. However, since JPT categories are extrinsic, the anticipated similarity may be masked. The only difference

Table 2: This table shows the total F_+ and F_- for two tests with the same CPD on equally sized data sets (2250 observations), drawn from the same populations. The only difference is the samples have different rCSs. Because the tests were run on data sets with different sized classes, the equivalence of the CPD's effectiveness is not obvious. It would be easy for an observer to erroneously conclude the CPDs were significantly different.

Relative class size	F_+	F_-
1 : 1	125	250
1 : 9	25	450

between the test outputs shown in Table 2 is one test set has a relative class size of 9 : 1 (rCS = 9) and the other a relative class size of 1 : 1 (rCS = 1). The CPD performs equally well in each test; however, rCS introduces a bias in the JPTs that makes the CPD performance equality difficult to recognize. When rCS = 1, there are twice as many F_- observations as F_+. However, when rCS = 9, the ratio between F_- and F_+ goes to 18 : 1! Without knowledge of the test sets used, an observer could well conclude that these were two significantly different CPDs, with significantly different applicability and/or efficacy. This is an obvious problem for researchers, thus significant effort has been applied to mitigate it; a selection of rCS invariant measures are available:

(i) the Youden index [15];

(ii) two related measures, diagnostic odds ratio (DOR) [16] and diagnostic power (DP) [17] (DP = $(\sqrt{3}/\pi)\log(\text{DOR})$;

(iii) the Matthews correlation coefficient (MCC) [18];

(iv) the receiver (or relative) operating characteristic[6] area under the curve (AUC) [19, 20];

(v) information theoretic measures such as the information coefficient (IC) [21, 22].

Youden was addressing the rCS's biasing effect in 1950; thus, the problem has been known for well over half a century, yet reports regarding mitigation are still entering the literature [9, 23]. In the works reviewed, consideration of end user problem environment was tightly constrained and the view of the data virtually unrestricted. We invert these criteria; first identifying problem interactions with rCS (a broad view of end user needs), then viewing the data such that it addresses the question posed by Jamain and Hand (a constrained view of the data).

rCS is generally confounding in the research environment; this is presumably also true for some end users. However, for other end users, rCS may be important for their problem. For these end users, basing decisions on rCS invariant measures may be misleading. Hence, we start by asking two questions:

(i) "Is rCS important for all end user CPD problem domains?"

(ii) "If not, what characteristics define when to incorporate rCS?"

Consider relating these two questions to a pair of real-world problems. A less effective CPD used with a rheumatoid arthritis test could lead to either more people than necessary being treated, or fewer. Likewise, a poorly selected intrusion detection boundary could cause an IT system to have excessive errors (false alarms or missed attacks).

In order to consider the two questions posed above, we will use a statistical nomenclature to describe a supervised CPD test bed. Viewed from a statistical perspective, observations on the dataset processed provide estimates of the underlying (class) population probabilities (rCS is the odds expression of that probability). Observations in the test system input are drawn from the specific populations (because the source class populations are known for each observation, a "ground truth" exists). The source population relative class sizes can be represented as a probability, for example, the probability that a randomly selected input will be a member of class A. This is the leading probability (P_{leading}), the probability before the inputs interact with the defined process.[7] In the examples stated, class A members would consist of RA-positive individuals and malicious information system activity. In the field, "ground truth" for any particular individual cannot be known prior to being processed (otherwise there would be no need for evaluation). However, in the test scenario being described, ground truth is known for each test set member. Since the source class is known for each CPD input element, input uncertainty does not affect any individual CPD output.

The balance of this paper is organized as follows. Section 2 considers relative class size. Section 3 describes the research protocol used in this study. Section 4 discusses efficacious measures for end users and considers existing summary statistics. Section 5 considers a cause of rCS invariance in measures and implications thereof. This is followed in Section 6 where we present two examples using the proposed format and tool. The main body of this paper closes with a summary of our findings and presents future work in Section 7. Four appendices with equation derivations and additional JPT normalization details wrap up the paper.

2. Relative Class Size

Abstractly described, test set elements interact with the defined process. This interaction "modifies" the elements (perhaps only by adding a tag indicating strength of the match with a model), leading to a test for class $\overline{A}$ membership. The probability that a randomly selected output will be detected as a member of class $\overline{A}$ is the subsequent probability ($P_{\text{subsequent}}$). $P_{\text{subsequent}}$ describes the state of the data stream after interacting with the defined process and is the combined result of the input mix (quantified as a probability ($P_{\text{leading}} = Y/N$) or an odds ratio (rCS = $\overline{Y}/Y$)) and the defined process. The defined process contributes its own uncertainty (P_{event}) to the observed output. Thus, the test system can be described by the equation $P_{\text{subsequent}} = f(P_{\text{leading}}, P_{\text{event}})$. The CDP test set model is illustrated in Figure 1.

Tying the test set model to the examples, $P_{\text{subsequent}}$ consists of the patient's RA diagnosis and the stream of

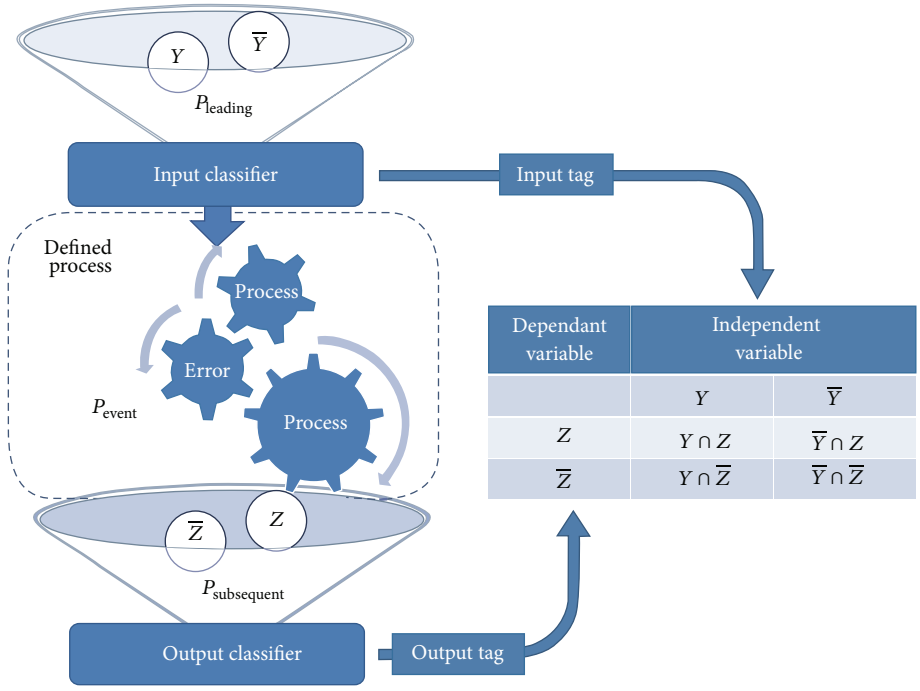


FIGURE 1: The test system "ground truth" inputs have a specific mix, representing the underlying probability for the system (P_{leading}). The test system outputs have a specific mix ($P_{\text{subsequent}}$), representing the interaction of the defined process and the inputs. The defined process contribution to the uncertainty observed in the output is represented by P_{event}. Often, the results are presented in JPTs.

intrusion detector classifications. P_{event} for the RA diagnosis consists of the strength of the match between the compound assayed and RA, test quality and the boundary used to determine class membership (diseased, not diseased). Similarly, P_{event} for the intrusion detection example consists of the appropriateness of the model that represents the malicious activity, the reliability of tags defining the activity, and the algorithm (or perhaps rule set) used to make malicious/non malicious determination.

In the CPD test system described, $\widehat{P}_{\text{leading}}$ is a characteristic of the input test dataset; hence, it is always fixed.[8,9] It is, in fact, related to rCS:

$$\widehat{P}_{\text{leading}} = 1 - \frac{\text{rCS}}{1 + \text{rCS}}. \quad (1)$$

We can now restate the original question within our framework: "are there problem domains where P_{leading} is important rather than confounding?"

One such situation could arise where individual results are significant only to the extent to which they contribute to a cumulative result. Consider setting intrusion detection boundaries. The end user is interested in limiting the impact of intrusions and intrusion prevention. The impact is cumulative, with each evaluation activity contributing. In this case, relative class size (expressed as P_{leading} or rCS) is important. If the end user were to base its boundary on P_{event} by using a rCS invariant measure (a measure that could not reflect the end user's estimate of their attack rate), there would likely be either excessive false alarm processing costs or excessive expenses due to missed attacks. Cases of this type, where each individual outcome contributes to a cumulative result, require knowledge of both P_{leading} and P_{event}.

Are there conditions in which P_{leading}, rather than being essential, might instead cause errors? We suggest that one such situation is when individual results are important and cumulative results are not. Consider a person tested for rheumatoid arthritis (RA). Depending upon the physician's office ordering the test, the frequency of RA_+s tested could vary considerably.[8] If each office set test boundaries to minimize their respective error rates, there would be a range of test scores that would be classified differently by different offices. Clearly, both diagnoses cannot be correct; a person cannot be simultaneously RA_+ and RA_-. In this case, considering the physician's rCS-based P_{leading} does not minimize the error for the patient.

With regard to rCS, we see that while basic research benefits from rCS invariant measures, these measures are not suitable for all end users; rCS invariance will be confounding for some end users. For these end users, any specific environment can have any of (literally) an infinite number of rCS values. Indeed, an end user's expected rCS can vary over time. Thus, a "one size fits all" solution will not be particularly efficacious. Our goal to provide for end user rCS needs, thus, resolves into two tasks:

(i) identify both rCS-sensitive and rCS-invariant measures that are efficacious for end users;

(ii) identify a means by which end users can tailor reported CPD results to reflect performance for their expected rCS.

3. Research Protocol

Although in many problem domains, populations tend to be normally distributed, this is not universal. In order to avoid limiting the applicability of our results, we use analytic procedures that are insensitive to distribution. To preserve generality, our analysis is strictly nonparametric; medians are used instead of means and quantiles are used instead of standard deviations. We also execute our tests with the Monte Carlo method, a nonparametric analytical tool often used when problem complexity (in our case, potential end user problem complexity) is not amenable to mathematical analysis.

CPD evaluation studies can be partitioned into two groups: those that use "real-world" data and those that use simulated data. Characterizing CPD evaluation measures requires observing how the measures respond as CPD output varies. Real-world data, such as those available in repositories, for example, the UCI Machine Learning Repository, provide the opportunity to test against a wide variety of complex data types [24]. However, observing the effect of incremental changes on real-world data is difficult at best. For our purpose, we use simulated CPD output. Although any distribution could be used, we assert normality when generating datasets. All data sets used in this study were generated such that the classes were normally distributed ($N(\overline{m}, \sigma^2)$; $\overline{m}$ is the distribution mean and σ is the standard deviation). The figures displayed were based on four hundred datasets consisting of two hundred thousand randomly drawn observations from two source populations: positive = $N(1.0, 0.0225)$ and negative = $N(2.0, 0.0625)$. Separate tests were run with datasets having rCSs of

$$2^0{:}1,\ 2^1{:}1,\ 2^2{:}1, \ldots,\ 2^{13}{:}1. \tag{2}$$

A total of 5,600 independent data sets were used in this study.

For each summary statistic evaluated, we observed how the reported metric was affected by rCS versus boundary versus metric output. The 3D results are presented as contour plots. Because the measure values are asymptotic to one (thus nonlinear), we use the median of the four hundred runs for each test case; means are not valid for non-linear scales. It is impractical to present confidence intervals on 3D data, but on the 2D graphs in Appendix D, the ninety percent confidence interval (90% CI) is displayed for select test series. To illustrate, the 90% CI is indicated by the vertical lines at each rCS tested in Figure 2; the horizontal line indicates the median.

This protocol provides the flexibility and repeatability necessary for analysis, yet abides by the constraints necessary for analysis of less tractable problem domains with difficult problem environments (e.g., complex CPD input and output distributions).

4. End User Efficacy

End users have two activities: CPD selection and CPD application. Regardless of any end user problem distinctions, these two activities address common interests:

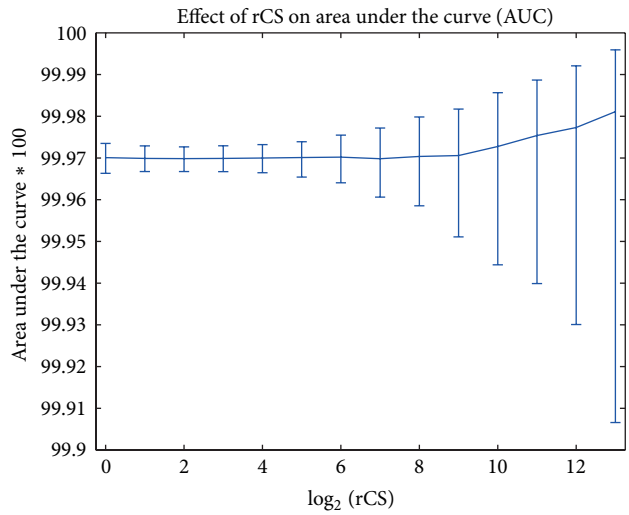


FIGURE 2: The vertical bars in this graph indicate the 90% confidence interval for measurements at each point observed. The horizontal line is the median.

(i) process application is concerned with the accuracy of the CPD for both possible outcomes:

  (1) "given that the test is positive, to what extent can the result be relied upon?"
  (2) "given that the test is negative, to what extent can the result be relied upon?"

  Mathematically, this can be expressed as a conditional probability, or a conditional odds. These values are monotonic, so difficult to use for optimum boundary identification;

(ii) process selection needs to choose the CPD with the best expected accuracy ("given the set of choices for CPD, which CPD will provide the best results and to what extent can its results be relied upon?"). Summary statistic output (based on the two monotonic measures) can inform end users for process selection.

The end user efficacious measures differ for the two rCS problem types; so they are further discussed in the following sections.

4.1. When rCS Is Important. Measure efficacy depends upon whether or not the impact on the end user is cumulative. The two CPD application questions can be expressed mathematically as

(i) "given that the test is positive, to what extent can the result be relied upon?" $\Omega(T_+ \mid Z) = T_+/F_+$;

(ii) "given that the test is negative, to what extent can the result be relied upon?" $\Omega(T_- \mid \overline{Z}) = T_-/F_-$.

Proportions, being asymptotic to one, are not ratio measures [25] and, thus, have limited utility. We use odds ratios instead.

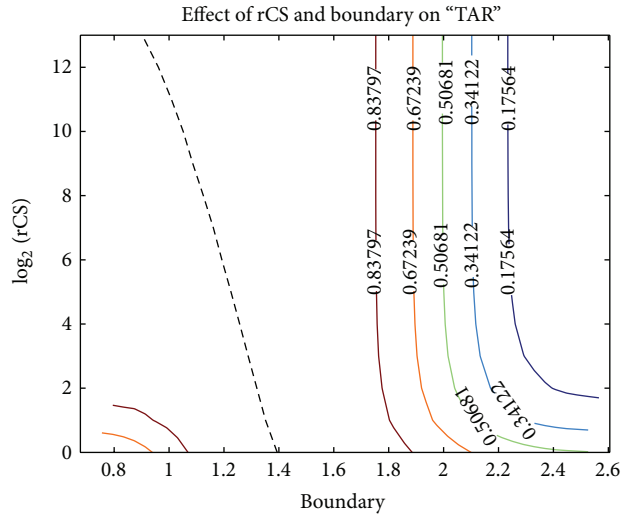


FIGURE 3: Being the sum of the observed correct classifications, TAR is a good measure for evaluating rCS sensitive CPDs. It is significant that the dashed line, indicating the optimum boundary, is not vertical; this shows that TAR is rCS sensitive.

The effect is cumulative (or additive); so the two conditional expressions, instead of being a measure suite, provide ancillary information. For CPD selection, end users will be interested in the proportion of the input stream that can be expected to be correct. Thus, a measure such as the total accuracy rate (TAR)

$$\text{TAR} = \frac{T_+ + T_-}{T_+ + T_- + F_+ + F_-} \tag{3}$$

represents the per element expected accuracy. TAR, being asymptotic to one, is not a ratio measure; therefore, averaging is not a valid operation. The total accuracy odds ratio (TOR) would be a better choice:

$$\text{TOR} = \frac{T_+ + T_-}{F_+ + F_-}. \tag{4}$$

Proportions, such as TAR and odds ratios such as TOR, are alternate expressions of the same CPD output. In fact, odds ratios can be transformed into proportions using (1).

Figure 3 shows TAR, as boundary and rCS vary. The TAR contour appears to vary little and be relatively constant over a wide boundary range. The optimum boundary (shown on the graph as the black dashed line) intersects the x-axis at around 1.45 and slopes toward 1.0. Additionally, the contour around the optimum boundary flattens as rCS increases. We can also see the optimum boundary and the reported accuracy rate both change as rCS varies. TAR is sensitive to rCS; thus, it is useful for problem domains where cumulative effects are important.

There are two other commonly seen rCS-sensitive summary statistics: F-score and Matthews correlation coefficient[9] (MCC). Another measure, information coefficient (IC); is becoming more prevalent, so we consider it as well.[10]

4.1.1. F-Score. F-score is the complement of a summary statistic proposed by van Rijsbergen [26]. The measure suite for F-score is recall and precision; van Rijsbergen's measure is based on information retrieval performance criteria put forth by Cleverdon [27]. Cleverdon's criteria address practitioner needs in information retrieval. Recall quantifies a CPD's completeness (the probability that the desired observations in the database are correctly identified). Precision quantifies the probability that undesired observations are mistakenly labeled as desired. For the information retrieval domain, these data[11] seem to be what end users need to know (F-score is now being seen in other problem domains.)

Recall and precision correspond to the conditional probabilities $P(T_+ \mid Y)$ and $P(T_+ \mid Z)$ (also known as "True positive rate" (TPR) and "positive predictive value" (PPV)). In the problem domain within which they were introduced (information retrieval), these measures quantify how well an CPD relates an object to a concept, such as selecting a document based on keywords. F-score is defined as

$$F_\beta = \frac{\left(1 + \beta^2\right)(\text{precision})(\text{recall})}{\left(\beta^2\right)(\text{precision} + \text{recall})}, \tag{5}$$

where β is the relative weight of precision and recall:

$$\beta = \frac{\text{importance of precision}}{\text{importance of recall}}. \tag{6}$$

If precision and recall have equal weights, then $F_\beta = F_1$, which is the harmonic mean of precision and recall:

$$F_1 = 2\frac{(\text{precision})(\text{recall})}{\text{precision} + \text{recall}}. \tag{7}$$

Since precision and recall are conditional probabilities, we can convert the F-score equation into JPT values. After substitution and rearranging terms,

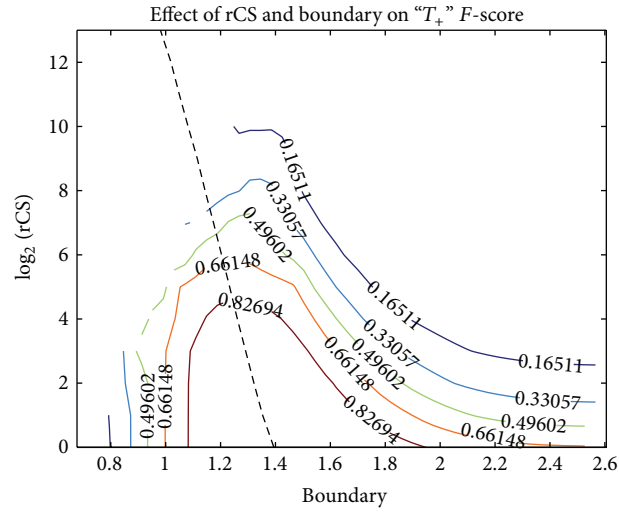
$$F_1 = \frac{T_+}{T_+ + F_+/2 + F_-/2}. \tag{8}$$

The derivation is provided in Appendix A. Notably, using the harmonic mean results in T_+, F_+ and F_- not being equally weighted in the denominator. While this may be suitable for the information retrieval domain and some others, it is hardly universal.

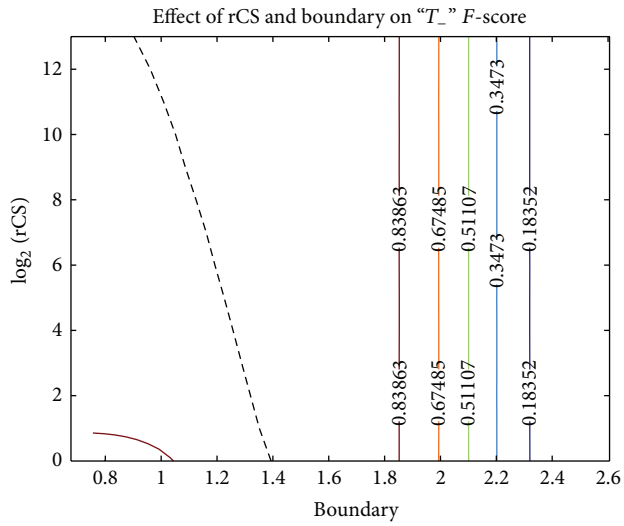
In contrast to TAR, F-score's rCS sensitivity varies, depending upon the class monitored. The effect can be seen in Figures 4(a) and 4(b). Interestingly, both TAR and F-score are well-accepted measures. This may indicate the existence of another CPD problem structure element. Analysis of this possibility is postponed for later consideration.

4.1.2. Matthews Correlation Coefficient. The Matthews correlation coefficient (MCC) is a more recent measure, introduced by Matthews [18]. MCC is the application of Pearsons correlation coefficient to CPD evaluation. In a subsequent classifier measure survey, Baldi et al. restated the measure in the form commonly seen today [21]:

$$\text{MCC} = \frac{(T_+ * T_-) - (F_+ * F_-)}{\sqrt{Y * \overline{Y} * Z * \overline{Z}}}. \tag{9}$$



(a) The relatively flat area around the optimum boundary (black dashed line in the graph) with low rCS suggests a low boundary sensitivity. The ridge follows the same optimum boundary as that of the total accuracy rate



(b) If, instead of selecting T_+, we select T_-, then F-score looks remarkably similar to TAR. Thus, F-score quantifies the categorical process's effect on a specific category

FIGURE 4: These graphs show that F-score, a summary statistic commonly used to compare CPD effectiveness, is rCS-sensitive. This is a desirable characteristic for problem domains such as information retrieval. In addition to rCS sensitivity, F-score is also sensitive to the target class.

Although not mentioned explicitly in Matthews' work, Baldi et al. note that the equation requires normalized distributions:

$$\frac{a - \overline{a}}{\sigma_A}, \quad \text{where } a \in \{A\}, \tag{10}$$

$A = \{a_1, a_2, a_3, \ldots, a_s\}$ is the class of the input dataset and $s = |A|$. $\overline{a}$ is the mean of A. $\overline{A}$ is treated in the same manner. One effect of distribution normalization is class size equalization: rCS = 1 (this is discussed in Section 5). The expressions used to generate normalized JPTs are shown in Table 3 the class size equalization is indicated by the ones in the "normalized totals" row. To demonstrate Matthews initial intent to use normalized distributions, we recalculated Matthews et al.'s original results using both actual and normalized JPTs. The values using normalized JPTs matched Matthews results; the values using actual JPT values varied from Matthews reported values by approximately a factor of twenty. Thus, MCC, when introduced, was intended to be calculated on normalized JPTs.

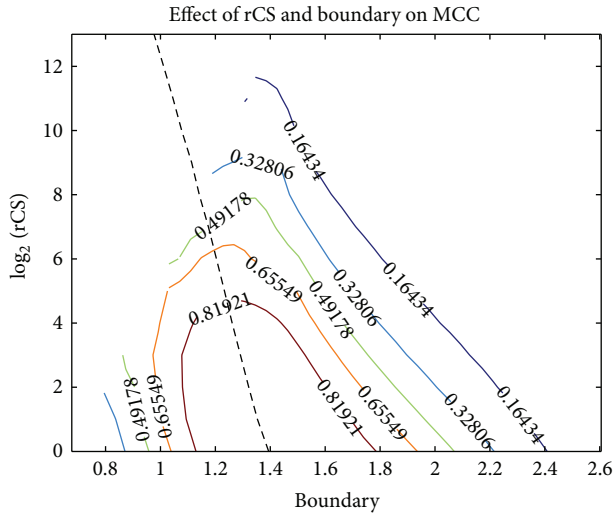
TABLE 3: The expressions in this JPT normalize the category values.

Actual target classification		Y	$\overline{Y}$	
Test result	Positive	T_+/Y	$F_+/\overline{Y}$	
	Negative	F_-/Y	$T_-/\overline{Y}$	
	Normalized totals	1	1	2

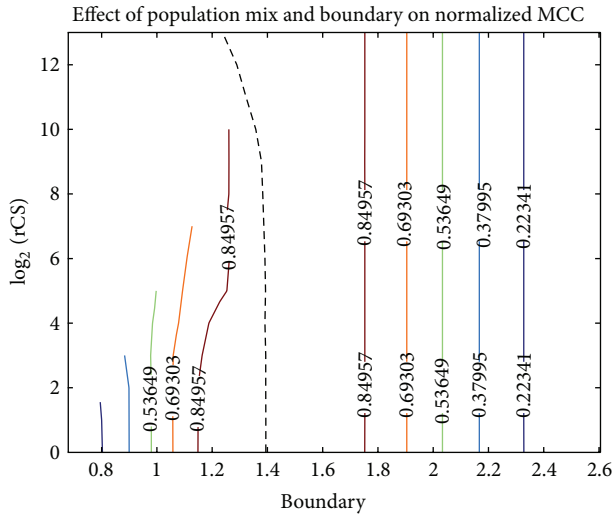
The JPT normalization prerequisite applied by Matthews et al. and noted by Baldi et al. seems to have been lost, although the belief that MCC is rCS-invariant persists [28–34].[12] As a consequence, Baldi et al.'s equation is sometimes applied without first normalizing the JPTs. In two of these reports, Cannon et al. [28] and Mirceva et al. [34] include JPTs. Upon recalculating their results, we determined that the values presented were based on nonnormalized JPTs. In both cases, there were substantial differences between the results on normalized and nonnormalized JPTs. In the Cannon et al. results, the difference affected not only the values, but also the process rankings. Using normalized JPTs, MOLPRINT, the process ranked last, moved into the upper fifty percent of processes tested. Having rankings of processes substantially change due to such changes could result in selection of a suboptimal process for use in real-world settings.

Figure 5 shows how normalization impacts the rCS sensitivity for MCC. Figure 5(a) shows MCC's response when the raw JPT values are used. The peak boundary (indicated on the graph by the dashed line) shifts as rCS varies and the value decreases as rCS increases (indicated by the sloping dashed line that intersects the contours). In contrast, Figure 5(b) shows that with JPT normalization, the peak boundary and calculated optimum MCC value are fixed (indicated by a vertical dashed line and contours that do not intersect the peak boundary). Exactly the same data sets were used for both graphs; the only difference is the presence or absence of JPT normalization.

Comparing MCC results is complicated by the fact that the published reports we surveyed did not identify whether or not the MCC values reported were on normalized JPTs or not. As seen in Figures 5(a) and 5(b), comparing results across tests where rCS is not normalized could lead to errors. Since rCS affects the optimum boundary when raw JPTs are used, a simple correction of reported values by recalculating MCC on normalized JPTs will most probably be for a suboptimal boundary; thus, the corrected MCC value will also be suboptimal.[13] In this section, we characterized MCC using nonnormalized data. Appendix C derives an rCS-invariant MCC expression.



(a) MCC exhibits rCS sensitivity



(b) On normalized JPTs, MCC exhibits rCS invariance. The increasing boundary curvature when rCS $> 2^6$ is a JPT normalization artifact explained in Appendix D

FIGURE 5: If MCC inputs are not normalized, it is rCS-sensitive.

Relative to end user interests, it is unclear what MCC quantifies and under what context the value will be relevant; MCC's end user efficacy is limited to optimum boundary identification.

MCC's complexity makes determining an underlying measure suite difficult. This detail will be addressed in the future.

4.1.3. Mutual Information Coefficient. Rost and Sander introduced an information-theory-based measure into the literature in 1993 [22]. It was subsequently included in a measure comparison by Baldi et al. [21]. Since then, it has gained some traction in biological literature [35–43] and has been seen in network management literature [44]. The measure is sometimes called the information coefficient or mutual information coefficient; we use the acronym IC.

As explained by Baldi et al., IC is the mutual information (I) normalized by the entropy in ground truth (H); I is the mutual information contained in ground truth regarding the test set $S(Y \cup \overline{Y})$ and the CPD prediction of ground truth, as contained in $Z \cup \overline{Z}$:

$$\mathrm{IC} = \frac{I\left(Y \cup \overline{Y}, Z \cup \overline{Z}\right)}{H\left(Y \cup \overline{Y}\right)}. \tag{11}$$

Expressing I and H in terms of JPT categories,

$$\begin{aligned} &I\left(Y \cup \overline{Y}, Z \cup \overline{Z}\right) \\ &\quad = -H\left(\frac{T_+}{N}, \frac{F_+}{N}, \frac{F_-}{N}, \frac{T_-}{N}\right) \\ &\qquad - \frac{T_+}{N}\log\left(|Y| * |Z|\right) - \frac{F_+}{N}\log\left(\left|\overline{Y}\right| * |Z|\right) \\ &\qquad - \frac{F_-}{N}\log\left(|Y| * \left|\overline{Z}\right|\right) - \frac{T_-}{N}\log\left(\left|\overline{Y}\right| * \left|\overline{Z}\right|\right), \end{aligned} \tag{12}$$

where

$$\begin{aligned} &H\left(\frac{T_+}{N}, \frac{F_+}{N}, \frac{F_-}{N}, \frac{T_-}{N}\right) \\ &\quad = -\frac{T_+}{N}\log\frac{T_+}{N} - \frac{F_+}{N}\log\frac{F_+}{N} - \frac{F_-}{N}\log\frac{F_-}{N} - \frac{T_-}{N}\log\frac{T_-}{N}. \end{aligned} \tag{13}$$

Information-theory-based measures are gaining traction in the literature [35–43]. Some of these reports indicate the belief that the measures are rCS-invariant [38, 40, 43]. Solis and Rackovsky [41] note that their particular information theoretic measure may not be rCS-invariant. The belief that information theoretic measures are rCS-invariant comes from the fact that information theory applies to probability density functions, which are always normalized (rCS = 1) [45, 46]. Unless JPTs are normalized prior to use, IC and related measures cannot be guaranteed to be rCS-invariant.

Like other measures, IC compares target CPD output to an CPD using random classification. However, it differs in that IC is based on the entropy existing in the test set and CDP output. If the input and output are the same, then IC = 1; if the output of the process is equivalent to that of random selection, then IC = 0. A side effect of IC's use of logs is increased computational complexity. All of the other measures evaluated have a complexity of $O(N)$, IC is $O(N^2)$. This may limit IC's utility for large data sets. IC's computational complexity did affect our analysis. Had we calculated IC on the two hundred thousand element test sets used for the other measures, it would have taken approximately six months. Consequently, we tested IC on twenty thousand element test sets. In Figure 6, we can see that the peak boundary shifts as rCS increases; thus, IC is not rCS-invariant. As with the other rCS-sensitive measures, JPT normalization can confer rCS invariance.

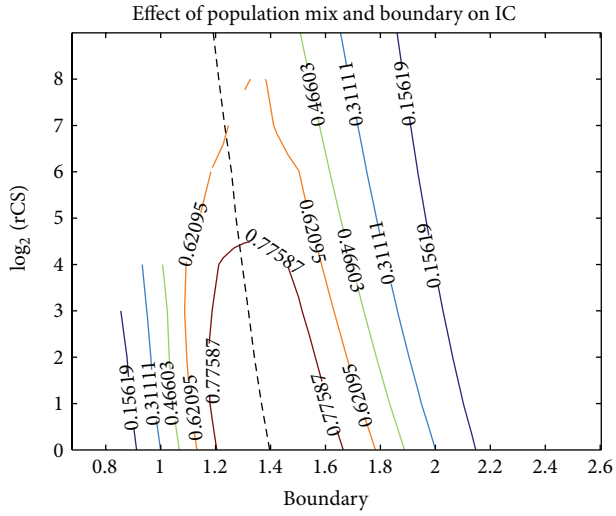


FIGURE 6: The sloped dotted line on the contour graph shows that IC is not rCS-invariant.

4.2. Findings regarding rCS Sensitive Problems

(i) Of the four rCS-sensitive summary statistics reviewed, TAR and F-score appear to be efficacious.

(ii) Because the reaction to rCS of TAR and F-score are opposite, this may indicate the existence of other elements in the CPD problem structure (we will address that in future work.)

(iii) MCC and IC, regardless of their apparent utility for researchers, do not seem to quantify information directly usable by end users.

This section has not covered how end users can take a single reported value and convert it into one applicable to their specific rCS environment. This will be discussed in Section 5.

4.3. When rCS Is Confounding. When rCS is confounding, in addition to quantifying end user issues, efficacious measures must be rCS-invariant. The following discussion will apply normalized JPT input when necessary.

The CPD application expressions for this problem type are normalized versions of those for rCS-sensitive problems.

"Given that the test is positive, to what extent can the result be relied upon?" Mathematically, this can be expressed as a conditional probability, or a conditional odds on normalized JPTs. For the reason mentioned in Section 4, we use the odds, normalized $\Omega(T_+ \mid Z) = T_+ \overline{Y}/F_+Y$.

"Given that the test is negative, to what extent can the result be relied upon?" The odds expression for this is normalized $\Omega(T_- \mid \overline{Z}) = T_-Y/F_- \overline{Y}$.

Test selection needs to choose the CPD with the best expected accuracy ("given that a result will be rendered, to what extent can the result be relied upon?"). The two CPD application questions provide operational information, but are also the basis for this noncumulative CPD problem selection decision. As such, they can be considered the measurement suite for the CPD selection decision. The CPD selection decision requires an end user efficacious summary statistic. The expected prediction accuracy (EPA) is the average of the two odds ratios identified in the previous paragraph. Each CPD event is independent and the conditional values are normalized. The special conditions that dictate applying either the geometrical mean (a compounding effect) or the harmonic mean (unequal set sizes) do not exist; so the arithmetic mean of the conditional odds on normalized JPTs is appropriate:

$$\text{EPA} = \frac{\left(T_+\overline{Y}\right)/\left(F_+Y\right) + \left(T_-Y\right)/\left(F_-\overline{Y}\right)}{2}. \tag{14}$$

To our knowledge, this end user summary statistic is not found in the literature. We apply this summary statistic in the meta-analysis in Section 6.

As noted in the introduction, rCS-invariant summary statistics are already in use. We review three commonly seen rCS-invariant summary statistics:

(i) the Youden index [15];

(ii) two related measures: diagnostic odds ratio (DOR) [16] and diagnostic power (DP) [17] (DP = $(\sqrt{3}/\pi)$ log(DOR));

(iii) the receiver (or relative) operating characteristic area under the curve (AUC) [19, 20].

Two other summary statistics, the Matthews correlation coefficient (MCC) [18] and mutual information coefficient (IC) [21] are commonly held to be rCS-invariant, but in fact are not. They were discussed in Section 4.1.

4.3.1. Youden Index. The Youden index (traditionally represented by J) was proposed in 1950 and is seen in medical diagnostic studies [15]. There are a number of expressions of J. The original is

$$J = \frac{1}{2}\left[\frac{T_+ - F_+}{T_+ + F_+} + \frac{T_- - F_-}{T_- + F_-}\right]. \tag{15}$$

Perhaps a more common representation is

$$J = \text{sensitivity} + \text{specificity} - 1, \tag{16}$$

where

$$\text{sensitivity} = \frac{T_+}{Y}, \qquad \text{specificity} = \frac{T_-}{\overline{Y}}. \tag{17}$$

Further, sensitivity is also known as the *true positive rate* (TPR) and specificity is the complement of the false positive rate; specificity = 1−FPR = $1-(F_+/\overline{Y})$. Hence, an even simpler (thus better, according to the minimum description length principal) definition would be

$$J = \text{TPR} - \text{FPR}. \tag{18}$$

In this form, the Youden index can be taken to be a summary statistic of the measure suite {TPR, FPR}.

J is special in that $J = 0$ indicates an CPD with an output equal to that of tossing a fair coin. $J = 1$ with a perfect CPD

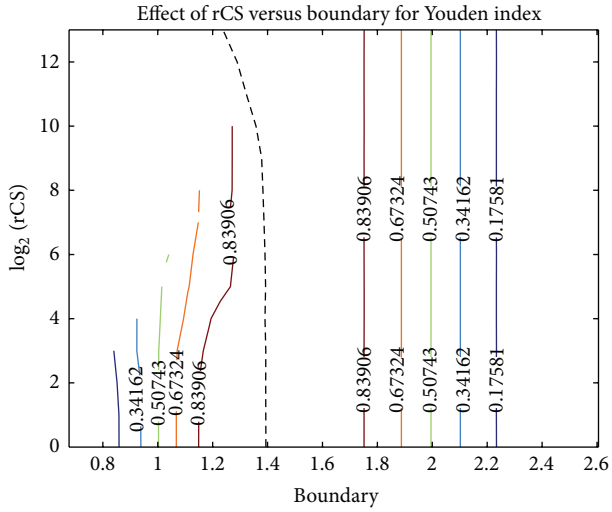


Figure 7: The Youden index has a very uniform shape and the optimum boundary lies along the peak of the Youden index ridge. This exhibits the expected rCS invariance.

and $J = -1$ for an CPD that misclassifies everything. As noted in their respective literature bases, J shares a characteristic with AUC, in that it is insensitive to rCS. On our source populations, the optimum boundary is approximately 1.4. This can be seen in Figure 7. There is an issue with end user efficacy, however. J quantifies the spread between the TPR and FPR. This information has little bearing on the "pretest" question posed at the beginning of this section.

4.3.2. Diagnostic Odds Ratio (DOR) and Discriminant Power (DP). Two related measures are the diagnostic odds ratio (DOR) [16] and discriminant power (DP) [17]. DOR is defined as

$$\mathrm{DOR} = \frac{T_+/F_-}{F_+/T_-}, \tag{19}$$

where T_+/F_- *is true positive odds* (TPO) and F_+/T_- is *false positive odds* (FPO). After simplification,

$$\mathrm{DOR} = \frac{T_+T_-}{F_+F_-}. \tag{20}$$

Discriminant power is defined as

$$\mathrm{DP} = \frac{\sqrt{3}}{\pi}\left(\log X + \log W\right), \tag{21}$$

where

$$X = \frac{\text{sensitivity}}{1-\text{sensitivity}}, \qquad Y = \frac{\text{specificity}}{1-\text{specificity}}. \tag{22}$$

Recasting the equation, we get

$$\mathrm{DP} = \frac{\sqrt{3}}{\pi}\log\left(\frac{T_+T_-}{F_+F_-}\right). \tag{23}$$

The derivation is provided in Appendix B. Comparing the two measures, we see that

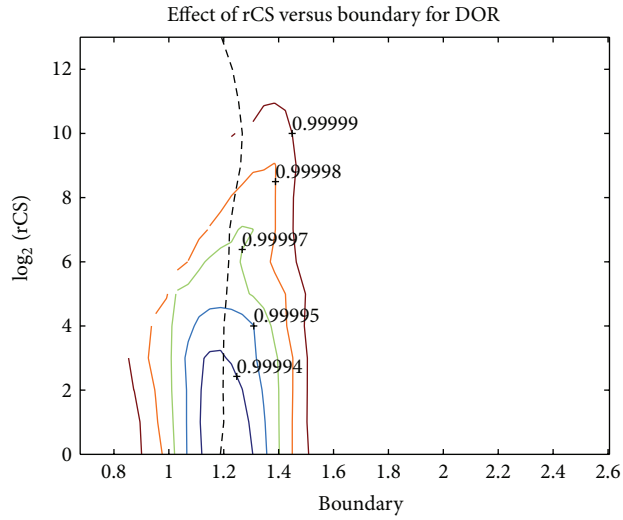
$$\mathrm{DP} = \frac{\sqrt{3}}{\pi}\log\left(\mathrm{DOR}\right). \tag{24}$$

DOR and DP are found in medical research. Interestingly, DP $= -\infty$ and DOR $= 0$ when either $T_+ = 0$ or $T_- = 0$, both need not equal zero. Similarly, DP $= \infty$ and DOR $= \infty$ when either $F_+ = 0$ or $F_- = 0$, both need not equal zero. Hence, an CPD can classify some observations correctly (total Accuracy > 0), yet have DP $= -\infty$ and DOR $= 0$. This is counterintuitive, since one would expect DP $= -\infty$ and DOR $= 0$ to indicate a totally ineffective CPD and DP $= \infty$ and DOR $= \infty$ to indicate a perfect CPD, rather than something in between. Since T_+ and T_- are (statistically) independent,[14] (as are F_+ and F_-), the DP and DOR could, in a probabilistic sense, be interpreted as the odds that, given two random observations, one will be classified T_+ and the other T_- (one will be classified F_+ and the other F_-). While the question seems similar, the fact that the DOR and DP optimum boundaries are different from the other inherently rCS invariant measures tested suggests that the two questions are significantly different. Perhaps this is because the DP and DOR treat the problem as a multiplicative function; we identify the problem as an additive function. This value would seem to be directly relevant in niche CPD scenarios, but not to general CPD problem types.

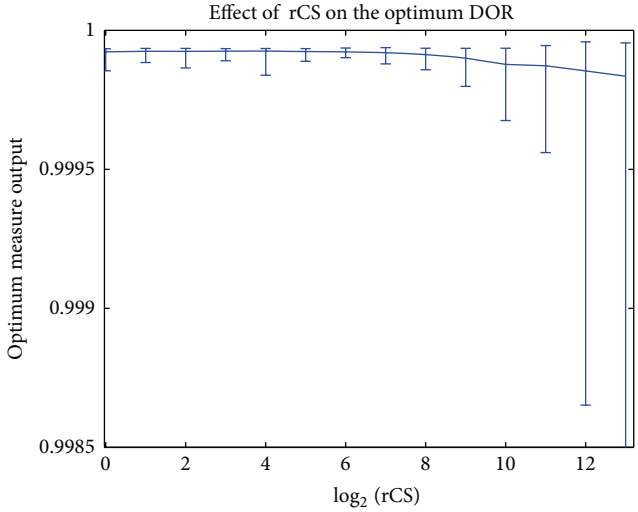
In medical studies, when the event tested for (T_+) has a low probability, DOR approximates relative risk: the rate at which the event was observed in group A versus the rate observed in group B. This is valuable information. However, when applied in the more general CPD domain, there is a problem. In any specific CPD task, the category of interest may not have a sufficiently low probability T_+; thus, the approximation may not always be acceptably close.

Unfortunately, DOR and DP have a challenging sensitivity to boundary; the optimum boundary is indicated by min(DOR) (or min(DP)). Thus, for any test run, the boundary with the smallest T_+T_- relative to F_+F_- gives the best accuracy. Not only is this counterintuitive, but also a potential error source. The problem originates from the fact that the greater the min(DOR) (or min(DP)), the better the results. Thus, if the boundary used to partition the test output is not at min(DOR) (or min(DP)), the results may appear better than they really are.[15] Most observations regarding DOR apply to DP as well. For example, DP $= 0$ when $T_+T_- = F_+F_-$.

One important characteristic of DOR and DP is that they are rCS invariant. An important difference between DOR/DP and the other rCS-invariant measures is that their optimum boundaries, although constant in our tests, are offset from the "minimum error boundary." These effects can be seen in Figure 8.[16] Since DOR and DP are minima, they follow a valley in the contour graph, instead of a ridge, Also contrary to the other measures, DOR decreases when the absolute class size effect becomes noticeable. This means that the contours are closed, instead of open as seen for the other measures. DOR's vertical optimum boundary line and constant value



(a) Contour graph of rCS versus boundary versus DOR value. Scaling makes the measure seem somewhat rCS sensitive. However, Figure 8(b) shows DOR is actually rCS invariant



(b) Graph of rCS versus DOR, with error bars

FIGURE 8: Instead of the optimum value being maxima, like the other measures evaluated, the optimum DOR value is a minimum. Hence, the contours show a valley instead of a ridge. Also contrary to the other measures, DOR decreases when the absolute class size effect becomes noticeable. This means that the contours are closed, instead of open like the others. DOR's vertical optimum boundary line and constant value (seen in Figure 8(b)) indicate that DOR (and hence, DP) is rCS-invariant. DOR/DP optimum boundaries (approx. 1.2) are offset from the optimum boundaries seen in the other rCS-invariant measures (approx. 1.4). DP, the log form of DOR, has the same characteristics as DOR.

(seen in Figure 8(b)) indicate that DOR (and hence, DP) is rCS-invariant. Because of this boundary bias, they may not be useful for selecting boundaries. For example, in our test environment, TAR at the common optimum boundary is 0.994, TAR at DOR optimum boundary is 0.958; the difference is significant at the 95% confidence level.

4.3.3. Receiver Operating Characteristic Area under the Curve (AUC). ROC has a solid history. Swets campaigned diligently to establish it as the evaluation criterion of choice [20, 47, 48]. The {TPR, FPR} measurement suite is the basis for the AUC summary statistic. The title originates from the fact that it is the area under a "ROC curve," a curve defined by false positive (FPR = $F_+/\overline{Y}$) and true positive (TPR = T_+/Y) rates. These values are calculated from JPTs of CPD output for a number of boundaries across the observed range, then graphed as the ROC curve [19, 49].[17] The literature describes the ROC curve (AUC) as being rCS-invariant as well as boundary-invariant. Because it is boundary-invariant, AUC is a popular tool in our present research environment. However, AUC has been criticized on theoretical terms recently [50, 51].

In contrast to the other summary statistics reviewed, AUC is generally accompanied by the ROC curve (indeed, the ROC curve may be presented without providing AUC). To a person skilled in the art, the ROC curve provides a great deal more information regarding CPD performance than does the single value AUC summary statistic[18] (this is, of course, true for any measure suite, since consolidation of multiple values into a single summary statistic value means that information is lost).

Compared to our end user focused criteria, ROC-AUC, being boundary-invariant, is not useful for boundary identification. Nor is it efficacious for end users.

There are numerous ROC-AUC variants [52, 53]. Vanderlooy and Hüllermeier determined in their comparison, that despite intuitive appeal, none of the variants confer any CPD selection improvement. From the end user perspective, since the underlying measure units remain the same, they all have the same limited efficacy.

Figure 9 shows the optimum boundary versus rCS for the proposed normalized PPV, NPV average and existing summary statistics, normalized MCC, Youden index, and DOR. As can be seen in the figure, DOR peaks at a different boundary than the other rCS-invariant measures tested, and (excepting DOR) the optimum boundary is relatively stable until rCS $> 2^6$, after which the detected optimum boundary starts dropping rapidly.[19] For our context, a key finding from Figure 9 is that not only does DOR have weak end user efficacy, it also should not be used to identify the optimum boundary.

4.4. Findings regarding rCS-Invariant End User Problems. For problems requiring rCS invariance, we find that

(i) end users need three values. For CPD selection, the expected total predictive accuracy (a summary statistic) EPA = $((T_+\overline{Y})/(F_+Y) + (T\_Y)/(F\_\overline{Y}))/2$ is important. When the CPD is used in the field, the summary statistic value has no meaning. Instead, end users need the information provided by the two measure suite elements; the positive predictive value odds ratio PPV = $(T_+\overline{Y})/(F_+Y)$ and the negative predictive value odds ratio NPV = $(T\_Y)/(F\_\overline{Y})$.

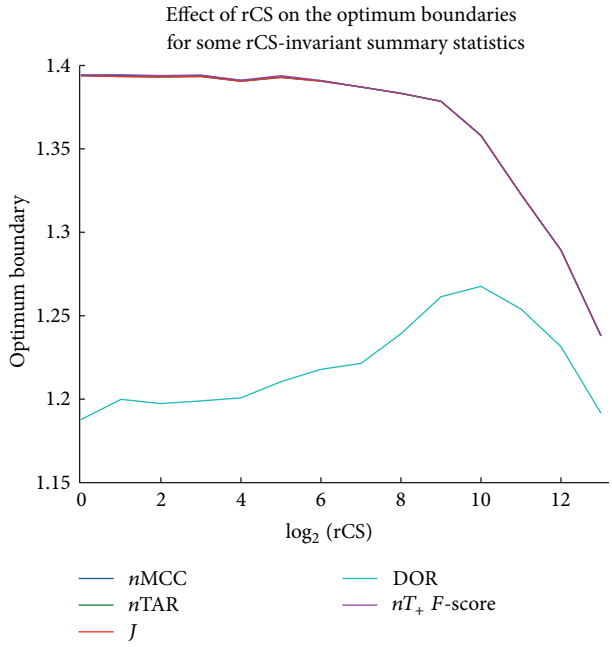


Figure 9: Other than DOR/DP, all of the rCS-invariant measures tested identified the same optimum boundary.

(ii) Although many commonly seen summary statistics can be used to identify the optimum boundary, as seen in this study, not all can (e.g., ROC-AUC and DOR). Figure 9 shows how the optimum boundary identified by DOR differs from that identified by the other summary statistics tested.

(iii) Of the rCS sensitive summary statistics evaluated, only the EPA output answers the end user's CPD selection question. The others may be useful for niche problems, but provide little useful information for the "common" end user.

5. JPT Normalization

In statistical circles, standardizing distributions is a well-established technique. One effect of standardization is that the area under the probability density function (pdf) equals 1. This simplifies pdf analysis, since the area of any segment of the area under the curve can be interpreted as the probability of an event occurring within that segment. Similarly, distribution standardization facilitates pdf comparisons. Since the CPD analysis domain considers processes with overlapping pdfs, it intersects with the pdf comparison domain, but is neither a superset nor a subset.[20] Where appropriate, distribution standardization is a useful tool.

In CPD analysis, distribution standardization takes the form of JPT normalization. Table 4 shows a JPT displaying "raw" data—actual category cardinality. After normalization, the class totals (bottom row in Table 5) are one. Thus, JPT normalization seems to be a cause for rCS-invariance in measures. As such, it provides a benefit to end users

Table 4: This JPT holds actual category counts.

Actual target classification		A	$\overline{A}$	
Test result	Positive	T_+	F_+	Z
	Negative	F_-	T_-	$\overline{Z}$
	Total	Y	$\overline{Y}$	N

Table 5: The values in this JPT have been normalized. Normalization results in equal class sizes (both total both equal one).

Actual target classification		A	$\overline{A}$	
Test result	Positive	T_+/Y	$F_+/\overline{Y}$	
	Negative	F_-/Y	$T_-/\overline{Y}$	
	Normalized total	1	1	2

with rCS invariant problems: any JPT-based CPD evaluation measure will have rCS invariant output, if the input JPTs are normalized. (Illustrated in Figure 10).

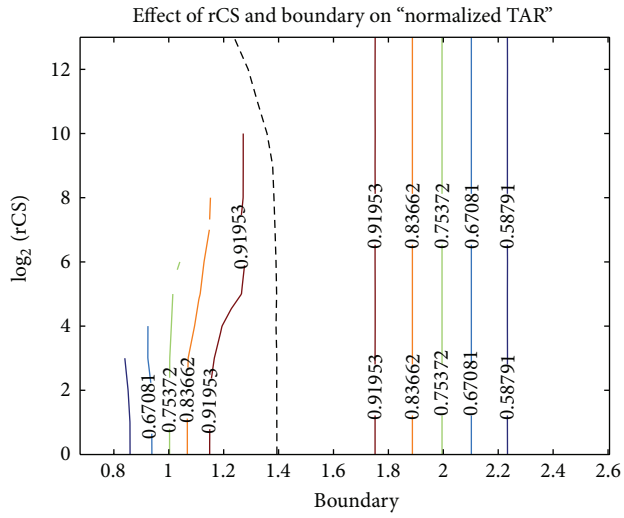
Although any measure can be rCS-invariant when the JPTs are normalized, some measures have emerged which have intrinsic rCS invariance. These inherently rCS invariant measures all have {TPV, FPV} (ratios that normalize the JPTs) as measure suites, thus rather than being counter examples, they provide empirical evidence that JPT normalization is the root cause for rCS invariance in measures; proof is beyond the scope of this paper. An overview of commonly seen rCS invariant measures is provided in Appendix D.

There is also a benefit for end users with rCS-sensitive problems. Statisticians use distribution standardization to mitigate rCS effects; however, the process is reversible. JPTs with rCS = 1 can be "tuned" to any desired rCS simply by multiplying one class by a constant c so that $c\overline{Y}/Y$ equals the desired value.[21] Thus, an end user with an rCS-sensitive problem can adjust reported results to fit their need. JPT tuning also allows end users to execute sensitivity analyses and estimate how the CPD will perform in their environment, over the expected rCS range. These insights are applied to a real-world problem in Section 6; a comparison of two RA diagnostic tests and an intrusion detection problem.

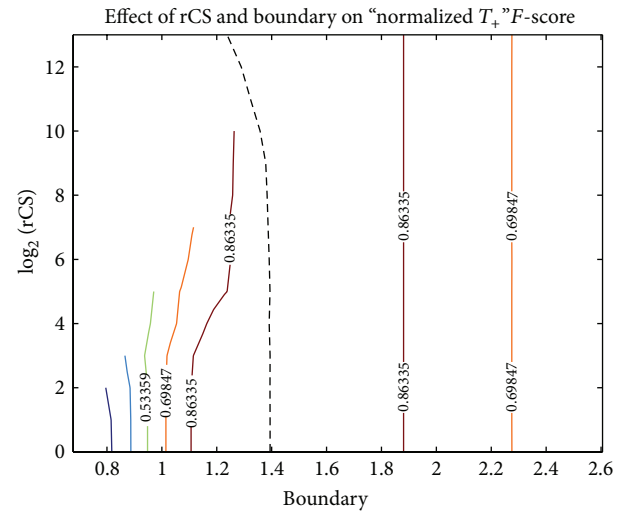
However, the optimum boundary is rCS-dependant; thus, the tool is not complete. To apply to all end users, results for all possible optimum boundaries would need to be provided.[22] This is impractical, if not impossible, for CPD test reports to include. As illustrated in Endnote 15, the tuned JPTs will indicate trends, but cannot be considered definitive. Nonetheless, JPT tuning extends JPT normalization in a way we have not previously seen in the literature and provides end users with a useful capability.

6. Examples

In this section, we use our proposed end user efficacious data analysis on two real-world problems, a meta-analysis[23] comparing two medical diagnostic tests for rheumatoid arthritis (RA) by [54] and data from a cyber security masquerading study by [55].



(a) Contour graph of normalized Accuracy rate. The reader may note that, other than the contour values, the graph is almost exactly the same as the Youden index graph



(b) Contour graph of normalized F-score. As with Youden index and normalized accuracy rate, the optimum boundary follows the "minimum error boundary"

FIGURE 10: The normalized accuracy rate and F-score seem to be relatively invariant to rCS. Not only is the value relatively constant, but the boundary stays constant as well.

6.1. A Meta-Analysis of Rheumatoid Arthritis Diagnostic Tests. The meta-analysis is quite thorough and accounts for many potential variations between studies. Three hundred and two relevant studies were found; eighty-six satisfied the rigorous inclusion criteria. The team concludes that one test is better than the other, however, does so without using a summary statistic. Our reanalysis adds the three recommended measures identified in Section 4.3.

The study uses two measures: positive likelihood ratio (LR_+) and negative likelihood ratio (LR_-). Given a typical test using supervised inputs (where ground truth is known), these two values are efficacious for researchers. They are less efficacious where end users have only the CPD output and ground truth is unknown. As we note in Section 4.3, PPV and NPV are more relevant for end users.

TABLE 6: Rheumatoid arthritis is a disease where both nontreatment and unnecessary treatment have negative consequences. Thus, knowing the overall predictive accuracy rate is useful information for a practitioner. These tables show the summary likelihood ratios originally reported and the corresponding normalized predictive accuracy odds ratios. Although the anti-CCP test is significantly better, its overall accuracy is not as great, nor is the negative predictive value as poor as one might believe, based simply on the likelihood ratio. (the parenthesized range in this and subsequent tables is the 95% confidence interval).

(a)

Test	Normalized odds ratio measures	
	LR_+	LR_-
Anti-CCP	12.46 (9.72–15.98)	0.36 (0.031–0.042)
RF	4.86 (3.95–5.97)	0.38 (0.33–0.44)

(b)

Test	Normalized odds ratio measures		
	PPV	NPV	EPA
Anti-CCP	13.4 (13.0–17.0)	2.88 (2.71–3.0)	8.14 (7.86–10.0)
RF	4.6 (4.25–5.0)	2.74 (2.63–2.87)	3.67 (3.44–3.93)

On pooled data, the LR_- differences between the tests were statistically insignificant. However, the LR_+ results were statistically significant. On the pooled data, the "anti-CCP"[24] positive test results were more frequently correct than the "RF"[25] diagnoses. The authors make one important point regarding rheumatoid arthritis treatment; it is harmful and costly to treat persons with false positive results. Hence, it is important to correctly diagnose negatives as well as positives: total expected predictive accuracy (EPA) is important for rheumatoid arthritis treatment. JPT normalization allows measurement of EPA that is rCS-invariant. Using normalized JPT data,

$$\text{EPA} = \frac{T_+/F_+ + T_-/F_-}{2}. \tag{25}$$

In our extension to Nishimura et al.'s report, we calculate the normalized EPA, PPV, and NPV on the pooled test data. Table 6 shows the original likelihood ratios reported by Nishimura et al. and EPA (the parenthesized range is the 95% confidence interval.[26]) LR_-s and PPV show similar values, but LR_+ is about one-sixth of the NPV; end users should be cautious when interpreting likelihood ratios. Comparing EPAs for each test and keeping in mind the end user context requires rCS invariance, the anti-CCP test correct diagnosis rate is a little more than twice the correct diagnoses rate of the RA test. This is true, even though, as can be seen in Table 7, the RF test actually more accurately detects RA's presence.

The authors note that "the better accuracy of anti-CCP antibody was mainly due to its higher specificity." In comparing the JPTs in Table 7, the anti-CCP pooled data

Table 7: These normalized JPTs of Nishimura et al.'s [54] pooled anti-CCP and RF test data were generated using their reported sensitivities and specificities. A person without RA is far less likely to be misdiagnosed than one with the disease, when the anti-CCP test is used.

(a)

		Actual RA condition	
		Diseased	Not diseased
Anti-CC test result	Positive	0.67 (0.65–0.68)	0.05 (0.04–0.06)
	Negative	0.33 (0.32–0.35)	0.95 (0.94–0.95)
	Total	1	1

(b)

		Actual RA condition	
		Diseased	Not diseased
RF test result	Positive	0.69 (0.68–0.7)	0.15 (0.14–0.16)
	Negative	0.31 (0.3–0.32)	0.85 (0.84–0.86)
	Total	1	1

Table 8: The above two JPTs have been "tuned" to a population where the diseased population is one hundred times the undiseased population. In this environment, the cumulative results will cause the anti-CCP test to appear to outperform the RF test.

(a)

		Actual RA condition	
		Diseased	Not diseased
Anti-CC test result	Positive	67	0.05
	Negative	33	0.95
	Totals	100	1
EPA		670 (850–650)	

(b)

		Actual RA condition	
		Diseased	Not diseased
RF test result	Positive	69	0.15
	Negative	31	0.85
	Totals	100	1
EPA		230 (250–212)	

summary in Table 7 shows that the anti-CCP test is actually less accurate in detecting diseased individuals—and at a statistically significant level (0.67 for anti-CCP is (statistically) significantly worse than the 0.69 reported for RF). We see that the anti-CCP actually detects RA less reliably than the RF test; the improvement is, in fact, entirely due to better specificity (correctly identifying nondiseased). In a case such as this, where each test is more accurate on one class, rather than one test being more accurate on both classes, it may not always be clear if there is any net diagnostic improvement. Normalized total predictive accuracy quantifies net diagnostic improvement in a way that may help clarify these issues.

This RA example is an rCS confounding type problem; in order to mitigate rCS bias, JPT normalization should be applied. We can now apply JPT tuning to illustrate how rCS can skew results; it is possible to estimate the cumulative test results that GPs and RA specialists will actually observe in their respective practices. The method actually "tunes" the JPTs; any desired rCS can be set.[27] A general practitioner may occasionally test for RA. Actual testing rates do not appear to be publicly available; so for computational simplicity, we assume the odds are one to one hundred that someone tested actually has the disease. Because of his/her specialty, a rheumatologist may have a new patient base that is highly skewed toward RA-positive. We assume a one hundred to one ratio. What total accuracy ratios will the two offices observe for the two tests? An JPT tuned to the rheumatologists' patient base is shown in Table 8, an JPT tuned to the GP's patient base is shown in Table 9. The EPA odds ratio observed by the rheumatologist would be anti-CCP: 670, RF: 250; the GP would observe total accuracy ratios of anti-CCP: 144, RF: 131, a statistically insignificant difference. Summary Table 10 shows that practices will have radically different experiences with the two tests, although the anti-CCP test is still best for the patient, regardless of the office.

6.2. A Cyber Security Masquerade Study. The cyber security problem domain is one where cumulative effects (e.g., processing false alarms (F_+)) are important. Consider an end user desiring to detect masquerading attacks, in which an attacker pretends to be an authorized user in order to gain access to a system. Determining the appropriate boundary for the detector is necessary in order to balance the effects of false alarms and missed attacks. This balance is subject to the relative volume of normal and masquerade system activity; thus, rCS is important; end users will want to incorporate rCS.

Schonlau et al. simulate a masquerade attack by capturing UNIX commands resulting from specific users, then inserting UNIX commands generated by another user into the original command stream. They compare a number of detection algorithms. The best performing was based on data compression. The Bayesian classifier [55] they used did not perform as well. We compare the two algorithms using the seven summary statistics discussed earlier.

Schonlau et al. use ROC curves to compare their various detection algorithms. From a research perspective, this is appropriate, since ROC is invariant to rCS and does not require boundary selection. However, as noted in Section 4.3.3, ROC provides limited information to end users. To illustrate the effect of using an inappropriate measure type, we reanalyze one of the user command streams with both rCS-sensitive and rCS-invariant measures. Table 11 shows the results.

For all seven summary statistics considered, the classifier with the higher value is better. Clearly, regardless of the measure, the compression algorithm outperforms the Bayesian classifier. The summary statistic values and associated optimum thresholds, however, vary widely.

What do we learn about the two classifiers from the measure values? The IC measures information content; MCC measures covariance. Youden index and DOR/DP quantify more esoteric characteristics. All four measure classifier

Table 9: The top two JPTs have been "tuned" to a population where the diseased population is one hundredth of the undiseased population. In this environment, the two tests appear statistically indistinguishable.

(a)

		Actual RA condition	
		Diseased	Not diseased
Anti-CCP	Positive	0.67	5
test result	Negative	0.33	95
	Totals	1	100
EPA	144 (136–150)		

(b)

		Actual RA condition	
		Diseased	Not diseased
RF test	Positive	0.69	15
Result	Negative	0.31	85
	Totals	1	100
EPA	137 (131–143)		

Table 10: This table shows the total accuracy ratios for RA tested populations of 100 : 1 (the rheumatologist) and 1 : 100 (the general practitioner).

Test	Patient bases (diseased : undiseased)	
	Rheumatologist (100 : 1)	General practitioner (1 : 100)
Anti-CCP	670 (850–650)	144 (136–150)
RF	230 (250–212)	137 (131–143)

performance relative to random selection using a fair coin, an issue particularly relevant to researchers, who generally consider a fair coin to be the most ineffective classifier. ROC-AUC, being rCS- and boundary-invariant, also has attractive characteristics for research. End users, however, are concerned about the net result, not distance from random selection. While each of these five measures quantify a characteristic related to the classifier performance characteristic of interest, none can be transformed into a value useful in the intrusion detection domain.

JPT tuning can help end users make more informed decisions. Schonlau et al.'s test sets consisted of one hundred blocks of concatenated UNIX commands. For "user 24," rCS = 3.7. "In the wild," one would expect rCS to be considerably smaller. For this example, we will assume that the end users expect rCS $\in$ [10 K, 100 K]. For the intrusion detection problem domain, TAR or F-score may provide the most information regarding end result. TAR includes both T_+ and T_-; F-score only includes T_+.[28] An IT system administrator may be most concerned about intrusion risk and detection overhead, thus not so concerned about T_-. If so, then F-score may be most relevant when comparing cyber security tools. Consider the ΩF-score of Schonlau et al.'s raw data in Table 12. The system administrator can tell that when rCS = 3.7, the correctly detected masquerade activity should be almost five times as frequent as errors; this is the system administrators greatest area of concern.

Table 11: These tables show the results for the compression-based classifier and the Bayesian classifier. The measures output on different scales and measure different characteristics; they cannot be directly compared. Because these are two different classifiers, the output ranges differ.

(a)

Compression classifier					
rCS-sensitive measures			rCS-invariant measures		
Measure	Value	Boundary	Measure	Value	Boundary
IC	0.528	0.800	Youden	0.791	0.200
TAR	0.930	0.800	DOR	6.78	0.200
MCC	0.786	0.600	ROC-AUC	0.851	NA
T_+ F-score	0.829	0.800			

(b)

Bayesian classifier					
rCS-sensitive measures			rCS-invariant measures		
Measure	Value	Boundary	Measure	Value	Boundary
IC	0.068	188	Youden	0.057	−228
TAR	0.620	638	DOR	0.206	638
MCC	0.053	−228	ROC-AUC	0.505	NA
T_+ F-score	0.543	−387			

Table 12: This table shows how JPT tuning can assist end users in estimating how an CPD will work in their environment. An executive looking at TOR will see that there are 25 correct classifications for every incorrect in the expected operating range. The IT system administrator looking at ΩF-score will see that there will be thousands of errors for every correct T_+. Their decisions regarding the usefulness of this CPD may differ.

rCS =	TOR	ΩF-score
1.0	7.8	7.1
3.7	13.3	4.9
1,000	19	0.042
10 K	25	0.0021
100 K	25	0.0004

A corporate executive might be concerned about the effect all four categories could have on the enterprise's performance; thus, TOR would be most relevant. Consider the TOR score of Schonlau et al.'s raw data in Table 12. The executive can tell that when rCS = 3.7, there will be over 13 correctly classified events for every misclassification. Based on these values, both persons might decide that performance is acceptable. JPT tuning, however, changes the picture considerably. The executive will see accuracy triple, but the system administrator will see a decrease in accuracy of more than three orders of magnitude. The executive and system administrator may now have different opinions.

Another problem with selecting an inappropriate summary statistic can be seen in Table 11. Not all measures have the same optimum threshold. An end user relying on an inappropriate summary statistic to determine a useful boundary for masquerade detection may be disappointed with their results.

After a classifier is selected, the two perspectives can lead to different system optimizations. When made available to end users, TAR and F-score values can help stakeholders such as executives and IT managers make more informed decisions. Since TAR/TOR and F-score/ΩF-score may have different optimum boundaries, practitioners and decision makers may benefit from having both values reported for each optimum boundary over an rCS range. That way, end users will have an appreciation of the tradeoff associated with selecting a particular solution.

7. Conclusion

This paper is a first step in identifying the structure of CPD problems faced by end users. Using that structure, we characterize how CPD evaluation measures are relevant to end users and identify end user relevant evaluation tools. To that end, we have defined rCS's importance to end user problems, identified measures that are efficacious for end users, and shown how JPT normalization and JPT tuning are useful for end user CPD evaluation.

Depending upon whether the end user is interested in the cumulative output or each individual CPD output, rCS is either an important factor or confounding. For maximum end user utility, research reports should include information efficacious for both problem types:

(i) for "rCS is confounding" problems, end users need a summary statistic, EPA $= ((T_+\overline{Y})/(F_+Y) + (T\_Y)/(F\_\overline{Y}))/2$ and the underlying measurement suite, PPV $= (T_+\overline{Y})/(F_+Y)$, NPV $= (T\_Y)/(F\_\overline{Y})$. All three values are based on normalized JPTs. If the values used are from normalized JPTs, then Y and $\overline{Y}$ both equal one, thus are unnecessary.

(ii) For "rCS is important" problems, end users must be able to tailor results to suit their individual rCS environments. We identify one appropriate summary statistic; the total accuracy odds ratio TOR $= (T_+ + T_-)/(F_+ + F_-)$. Another, F-score, is already in use: $F_\beta = ((1 + \beta^2)(\text{precision})(\text{recall}))/((\beta^2)(\text{precision} + \text{recall}))$, where β is the relative weight of precision and recall: β is the importance of precision relative to the importance of recall. End users can apply JPT tuning to tailor results for their environment. To do so, they will require the base JPT values ($\{T_+, F_+, F_-, T_-\}$).

Consolidating these findings, we propose that end users will be better served if research reports include PPV, NPV, EPA (or F_β, if it is prevalent in the domain), and the four normalized base JPT values.

Future work will continue to develop a CPD problem structure. The disparity between TAR and F-score suggests that at least one more characteristic exists. Also, without compensating for the effect of the shift in optimum boundary, JPT tuning does not fully address the end user's need to tailor research results. We will be considering means of addressing that deficiency.

Appendices

A. Restating F_1 in Terms of JPT Values

As defined,

$$F_1 = 2\frac{(\text{precision})\,(\text{recall})}{\text{precision} + \text{recall}}, \tag{A.1}$$

where

$$\begin{aligned} \text{precision} &= \frac{T_+}{T_+ + F_-}, \\ \text{recall} &= \frac{T_+}{T_+ + F_+}. \end{aligned} \tag{A.2}$$

Substituting, we have

$$F_1 = 2\frac{\left(T_+/\left(T_+ + F_-\right)\right)\left(T_+/\left(T_+ + F_+\right)\right)}{\left(T_+/\left(T_+ + F_-\right)\right) + \left(T_+/\left(T_+ + F_+\right)\right)}. \tag{A.3}$$

Multiplying and creating common denominators,

$$F_1 = \frac{2T_+^2/\left(\left(T_+ + F_-\right)\left(T_+ + F_+\right)\right)}{\left(T_+\left(T_+ + F_+\right) + T_+\left(T_+ + F_-\right)\right)/\left(\left(T_+ + F_-\right)\left(T_+ + F_+\right)\right)}. \tag{A.4}$$

Multiplying numerator and denominator by $(T_+ + F_-)(T_+ + F_+)/T_+$ leaves

$$F_1 = \frac{2T_+}{2T_+ + F_+ + F_-} = \frac{T_+}{T_+\left(F_+/2\right) + \left(F_-/2\right)}. \tag{A.5}$$

B. Restating DP in Terms of JPT Values

$$\text{DP} = \frac{\sqrt{3}}{\pi}\left(\log X + \log W\right), \tag{B.1}$$

where

$$\begin{aligned} X &= \frac{\text{sensitivity}}{1 - \text{sensitivity}}, \\ Y &= \frac{\text{specificity}}{1 - \text{specificity}}, \\ \text{sensitivity} &= \frac{T_+}{Y}, \\ 1 - \text{sensitivity} &= \frac{F_-}{Y}, \\ \text{specificity} &= \frac{T_-}{\overline{Y}}, \\ 1 - \text{specificity} &= \frac{F_+}{\overline{Y}}. \end{aligned} \tag{B.2}$$

Combining the logs, we get

$$\text{DP} = \frac{\sqrt{3}}{\pi}\left(\log\left(XY\right)\right). \tag{B.3}$$

Then, substituting for X and Y,

$$\mathrm{DP} = \frac{\sqrt{3}}{\pi}\left(\log\left(\frac{\text{sensitivity}}{1-\text{sensitivity}}\frac{\text{specificity}}{1-\text{specificity}}\right)\right). \quad \text{(B.4)}$$

Substituting for sensitivity and specificity,

$$\mathrm{DP} = \frac{\sqrt{3}}{\pi}\log\left(\frac{T_+/Y}{F_-/Y}\frac{T_-/\overline{Y}}{F_+/\overline{Y}}\right). \quad \text{(B.5)}$$

Multiplying top and bottom by $Y\overline{Y}$, we are left with

$$\mathrm{DP} = \frac{\sqrt{3}}{\pi}\log\left(\frac{T_+T_-}{F_-F_+}\right). \quad \text{(B.6)}$$

C. Derivation of Normalized MCC Equation

An important side note is that MCC, as commonly calculated,

$$\mathrm{MCC} = \frac{(T_+T_-)-(F_+F_-)}{\sqrt{Y\overline{Y}Z\overline{Z}}} \quad \text{(C.1)}$$

is not rCS-invariant as is sometimes reported [28–34]; it must use normalized JPT values (as in Table 13).

Substituting the normalized JPT values in (C.1) and collecting terms, the rCS-invariant MCC is

$$\text{normalized MCC} = \frac{(T_+T_-)-(F_+F_-)}{\sqrt{Y\overline{Y}\left(\overline{Y}T_+ + YF_+\right)\left(YT_- + \overline{Y}F_+\right)}}. \quad \text{(C.2)}$$

Equation (C.2) can be used in lieu of normalizing JPTs prior to calculating MCC.

D. Measures with Intrinsic rCS Invariance

Although the AUC, Youden index and DOR/DP are distinctly different measures; they all have one key similarity: normalized input. The AUC and Youden index both are (TPR, FPR), and since TPR and FPR are conditional probabilities $P(T_+ \mid Y)$ and $P(F_+ \mid \overline{Y})$, likewise, TNR and FNR are conditional probabilities $P(T_- \mid Y)$ and $P(F_- \mid \overline{Y})$. If, in the JPT, we replace T_+ by TPR, F_+ by FPR, T_- by TNR, F_- by FNR, then the marginal totals Y and $\overline{Y}$ are replaced by 1 s and N becomes 2. This is shown in Table 15. Since the two marginal totals representing class size are equal, this process compensates for rCS: the CPD output JPTs have been normalized. In this paper, calculations and discussion using the rCS invariant JPT form shown in Table 14 will refer to "normalized" versions. Any discussions not referring to "normalization" are of measures using the "raw" JPT form as presented in Section 3, Table 1.

Regardless of the actual test set rCSs, the input values for the AUC and Youden index incorporate JPT normalization. Although not as evident, this is also true for DOR and DP. Any JPT can be defined in terms of the TPR and FPR. This is illustrated in Table 14. Using Table 14 definitions,

$$\mathrm{DOR} = \frac{\left(c_Y\mathrm{TPR}\right)\left(c_{\overline{Y}}\left(1-\mathrm{FPR}\right)\right)}{\left(c_Y\left(1-\mathrm{TPR}\right)\left(c_{\overline{Y}}\mathrm{FPR}\right)\right)}. \quad \text{(D.1)}$$

Table 13: The values in this JPT have been normalized.

Actual target classification		Y	$\overline{Y}$	
Test result	Positive	T_+/Y	$F_+/\overline{Y}$	
	Negative	F_-/Y	$T_-/\overline{Y}$	
	Normalized totals	1	1	2

Table 14: JPTs can be defined in terms of the TPR and FPR. c_Y and $c_{\overline{Y}}$ are the class sizes in the test set.

		Source population		
		Y	$\overline{Y}$	Totals ↓
Test result	Positive	$c_Y * \mathrm{TPR}$	$c_{\overline{Y}} * \mathrm{FPR}$	Z
	Negative	$c_Y * (1-\mathrm{TPR})$	$c_{\overline{Y}} * (1-\mathrm{FPR})$	$\overline{Z}$
	Totals	c_Y	$c_{\overline{Y}}$	N

Table 15: A normalized JPT has class sizes adjusted to one. The four classification categories are expressed as proportions of the test set class of which they are actually members.

		Source population		
		Y	$\overline{Y}$	
Test result	Positive	TPR	FPR	
	Negative	FPR = 1 − TPR	TNR = 1 − FPR	
	Totals	1	1	2

After simplification,

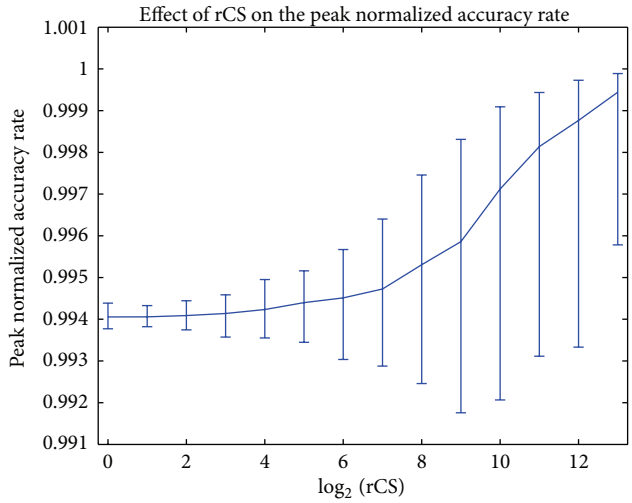
$$\mathrm{DOR} = \frac{\mathrm{TPR}\,(1-\mathrm{FPR})}{\mathrm{FPR}\,(1-\mathrm{TPR})}. \quad \text{(D.2)}$$

Thus, we find that DOR and DP are based on normalized JPTs as well.

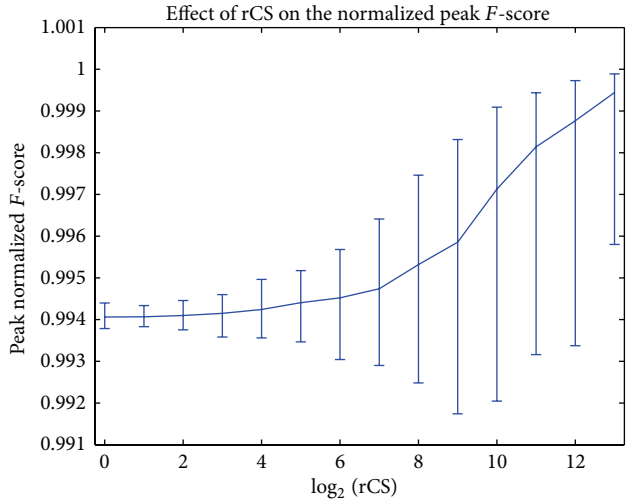
From the literature, we see that MCC is rCS-invariant when calculated on normalized JPTs. Presumably, other rCS-sensitive measures will be rCS-invariant when calculated on normalized JPTs as well. We tested this hypothesis by calculating accuracy, F-score, and MCC values on normalized versions. Figure 11 displays the peak Accuracy rate and F-score on normalized JPTs and compares them to the output of the established rCS-invariant measures, AUC, DOR,[29] and Youden index (DP, being just a log expression of DOR was left out.) The graphs are provided solely to compare their response to rCS. Any conclusions from Figure 11 beyond that must be made with care.

Figure 11 brings out some interesting points.

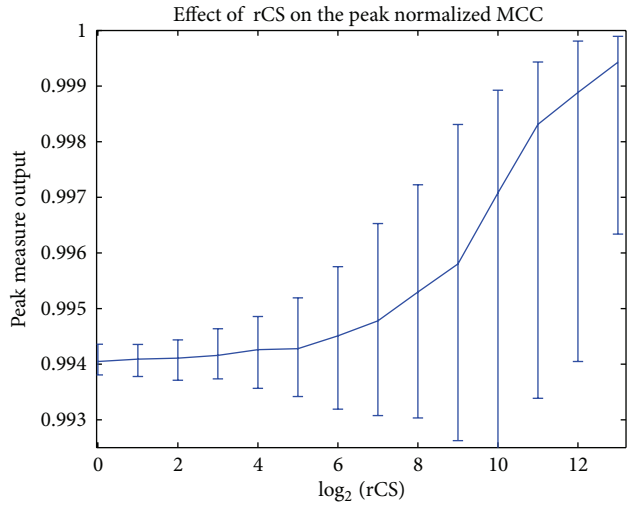
(i) Confidence interval response to rCS seems to fall into two categories. All of the normalized measures (including AUC, Youden index, DOR, and DP) have relatively stable CIs below rCS = 2^6. Above rCS = 2^6, there is an observable trend away from the stable value. This is due to a well-known issue with absolute sample size related to the strong law of large numbers. In our tests, the problem became statistically significant when the smaller sample had less than four hundred members.



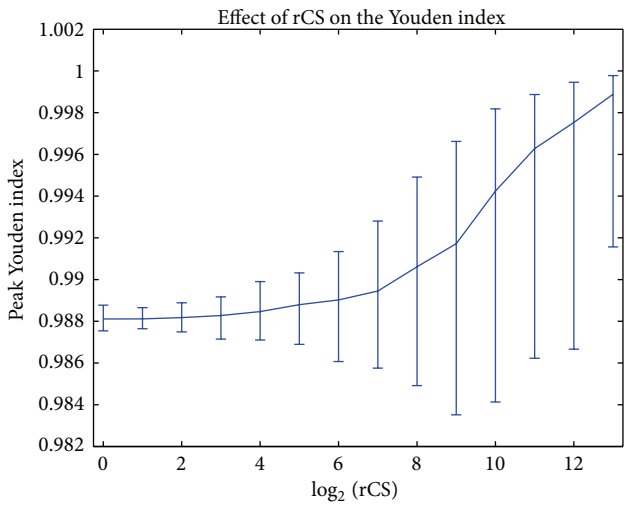
(a) Peak normalized accuracy rate (90% CI). It strongly resembles the Figures 11(b), 11(c), and 11(d)



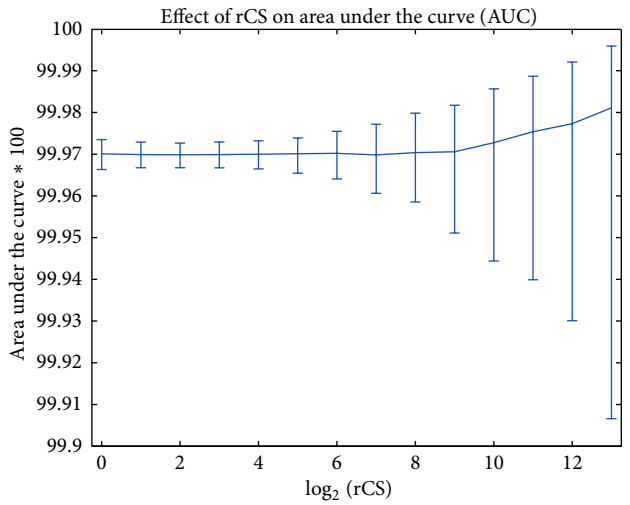
(b) Peak normalized F-score (90% CI). when $\beta = 1$, it strongly resembles Figures 11(a), 11(c) and 11(d)



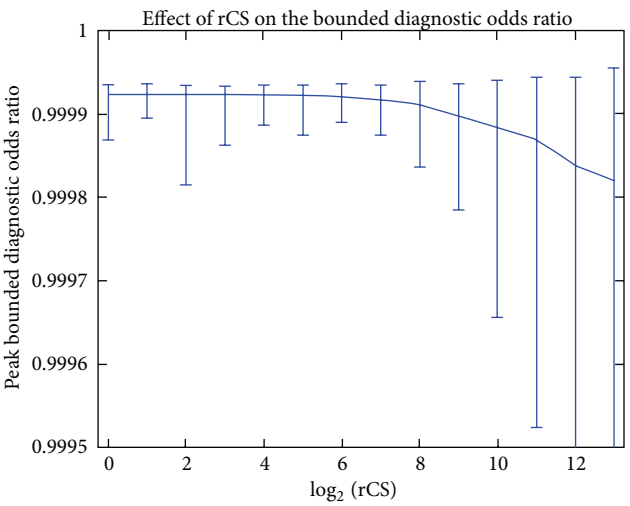
(c) Peak normalized MCC (90% CI): it strongly resembles Figures 11(a), 11(b) and 11(d)



(d) Peak Youden index (90% CI). This measure turns out to be related to the normalized accuracy rate



(e) Peak AUC (90% CI). It appears somewhat less sensitive to absolute sample size



(f) Best DOR (90% CI). As rCS invariance weakens, the DOR value drops

Figure 11: All of the normalized summary statistics exhibit rCS invariance. All of the lines vary from the horizontal, indicating that rCS invariance weakens when rCS $> 2^6$. This is a well-known absolute sample size issue. In our tests, the problem became statistically significant when $|A| < 400$.

For normalized accuracy rate, Youden index, normalized F-score, and normalized MCC, the 90% confidence interval generally increases as rCS increases. Analyzing the CIs is difficult because (all but DOR) are measured on scales with an upper bound, their scales are not linear. The CI changes observed, however, are consistent with expectations. In general, as the positive class size decreases, normalization magnifies any changes in T_+ and F_- far more than normalization of the negative class makes offsetting reductions (the positive class decreases by a factor of 2^{14}, while the negative class increases by a factor of less than 2^1).

(ii) Measure families have been found in the summary statistics evaluated. As discussed earlier, DOR and DP are related. The test also reveals a similarity between the normalized accuracy rate and Youden index:

$$\begin{aligned} \text{Youden index} &= \text{TPR} - \text{FPR} \\ \text{norm accuracy rate} &= \frac{\text{TPR} + 1 - \text{FPR}}{2} \end{aligned} \tag{D.3}$$

so that

$$\text{norm accuracy rate} = \frac{\text{Youden Index} + 1}{2}. \tag{D.4}$$

Thus, we see that normalized accuracy rate and Youden index are related.

(iii) JPT normalization can inflate reported process accuracy. Each graph in Figure 11 exhibits rCS stability when rCS $< 2^6$. However, when rCS $> 2^6$, rCS invariance seems to weaken. This turns out to be a function of the absolute size of the smaller class and is a consequence of the strong law of large numbers. As class sample size decreases, its representation of the source population decreases. The problem is that as sample size decreases, distribution tails lose their definition. When a sample size is magnified by JPT normalization, the undefined tails do not reappear, thus causing the sample to represent a source population with a smaller variance. This means the class overlap is underrepresented. Since process accuracy is inversely related to class overlap, a reduction in estimated class overlap will result in process accuracy overestimation. In our tests, the difference became statistically significant when sample sizes fell below four hundred members.

Violating the strong law of large numbers also affects the optimum boundary. As the apparent source population variance decreases, the boundary shifts toward that class. This can be seen in all of the contour graphs. In order to increase rCS, our protocol decreases $|A|$. A has the lower mean; thus, as rCS increases, the calculated optimum boundary starts shifting toward μ_A. In our tests, when rCS $> 2^6$, the shift becomes statistically significant.

Acknowledgments

The authors are thankful for the insightful comments in the early days of this work by Dr. Lynda Ballou, Mathematics Department, NM, USA Institute of Mining and Technology, Socorro, New Mexico. They are also indebted to Dr. Andrew Barnes, Applied Statistics Lab, General Electric Global Research, Niskayuna, NY, USA for taking the time to discuss this work. The journal reviewer's comments strengthened the paper considerably.

Endnotes

1. There are two levels of tool development. If development is "basic research," then evaluation is application-agnostic. If it is "applied research," then the focus is application-specific and evaluation needs tend to align with practitioner needs. For the purpose of this discussion, researchers do basic research; end users consist of practitioners and applied researchers.
2. CPD are a subset of the more general group of categorical problems. Our investigations apply to both.
3. The measure suite members quantify some particular aspect of CPD performance, thus providing greater CPD performance detail. These measure suite elements tend to be monotonic; thus, they are difficult to use individually to quantify overall CPD performance.
4. If the summary statistic only generates a single value, it is by definition B^*. B^* can also be a range. If there are multiple B^*s, they need not be continuous.
5. For example, measures used for CPD evaluation have been tested for invariance to various JPT perturbations [6]. Sokolova and Lapalme claim to be the first to comprehensively assess invariance to JPT perturbations; no boundary invariance assessments are known and only one summary statistic, AUC, claims boundary invariance. This study observes boundary effect on the metrics, but does not look for a basis for boundary invariance. This is left for future work.
6. The receiver operating characteristic originated in signal theory and gained its name from that problem domain. ROC, however, is now commonly used to analyze categorical data represented in JPTs. Although "receiver operating characteristic" may be the most commonly seen label in the literature, "relative operating characteristic," being less domain specific, has been proposed as a more appropriate name.
7. End user knowledge regarding actual input population probabilities for their environment may vary. As pointed out later, such information may be either important or confounding.
8. The example in Section 6 presents a case with two physician's offices. One was a general practitioner, where patients tested for RA had an rCS of 0.01 (RA_+/RA_-). The other was a rheumatologist. In that office, patients tested for RA had an rCS of 100 (RA_+/RA_-).

9. MCC is often touted as being rCS-invariant. This, however, is only true in a special case. This and related implications are discussed in Section 5.
10. IC is also considered to be rCS-invariant. As with MCC, this is only true in a special case.
11. Or their complements, the probability that desired observations are incorrectly identified and the probability that desired observations are mistakenly labeled as undesired.
12. A Google scholar search for "Matthews correlation coefficient" turned up well over one thousand articles. The publications cited are but a small sample.
13. Using the results shown in Figures 5(a) and 5(b) as an example, if a test was run on a sample with rCS = 2^8 on raw JPTs, the MCC $\simeq$ 0.33 and $B^* \simeq 1.1$. Recalculating MCC for the normalized JPT observed at rCS = 2^8 and $B^* \simeq 1.1$, results in MCC < 0.69. However, the actual peak is MCC > 0.85.
14. Independence is a highly overloaded term. In this context, it means that any change to T_+ will not affect T_-.
15. DOR and DP are seen frequently in medical studies. In this problem domain, the inappropriate boundary risk may not always be present. The risk would exist in a study of heart attacks versus cholesterol levels; cholesterol level is a continuous variable. However, in a study of heart attacks versus family history, family history could be binary (a close relative died/did not die). In this type of test, boundary sensitivity is not an issue; care must be taken, however, in test design. Just by changing the test to a count ("how many close relatives died/did not die," for instance) causes the problem to reappear.
16. DP and DOR are measured on different scales than the other summary statistics. In order to facilitate comparison, DOR was converted from an "odds ratio" type measure (bounded by $[0, \infty]$) to a "probability" type measure (bounded by $[0, 1]$). The relation between the two forms is

$$\text{probability measure} = 1 - \frac{1}{\text{odds measure} + 1}. \quad \text{(D.5)}$$

17. There is a similar measurement suite the "detection error tradeoff" (DET) [56]. DET plots the missed detection rate instead of the correct detection rate on the y axis. Since the two values are each other's complement, comments herein regarding ROC apply equally to DET. Interestingly, DET is plotted using log scales. This is a real challenge for measures with a lower bound of zero.
18. Since all of the inherently rCS-invariant measures studies have {TPR, FPR} as measurement suites, the ROC curve could be presented for each of them as well.
19. We noticed a similar effect on the measure's values. The values started becoming overly optimistic (once again, excepting DOR, the values of which dropped). The cause turned out to be a well-known issue with absolute sample size. The effect became significant when class A's size fell below 400 elements.
20. Pdf standardization is confounding when evaluating a problem where rCS is important; thus, standardization is not appropriate for all problems. The set of all CPD problems is greater than the set of problems where pdf standardization is useful. Likewise, the set of all pdf comparisons includes problems with other than overlapping (or potentially overlapping) probability distribution functions, hence, the CPD problem domain intersects with the pdf comparison domain, but is neither a superset nor a subset.
21. This expression does not require that rCS = 1 initially. With the exception of Y or $\overline{Y}$ equaling zero, any JPT can be transformed (tuned) from one rCS to another.
22. There may be a solution to this deficiency; we will investigate this in future work.
23. As nicely summarized by [57], meta-analysis is a statistical technique for combining the findings from independent studies. Meta-analysis is most often used to assess the clinical effectiveness of healthcare interventions; it does this by combining data from two or more randomized control trials. Meta-analysis of trials provides a precise estimate of treatment effect, giving due weight to the size of the different studies included. The validity of the meta-analysis depends on the quality of the systematic review on which it is based. Good meta-analyses aim for complete coverage of all relevant studies, look for the presence of heterogeneity, and explore the robustness of the main findings using sensitivity analysis.
24. Anti-CCP refers to an assay using cyclic citrullinated peptide (CCP) to detect the anti-CCP antibody.
25. RF is an initialism referring to rheumatoid factor, an antibody used as a marker for RA.
26. On a single tailed test as used here, only one bound is relevant; thus, the bound indicates a 97.5% confidence.
27. In a CPD setting where rCS is important, JPT tuning enables a capability previously unavailable: sensitivity analysis. For the practitioner, this means that CPD performance can be evaluated for the expected rCS. Moreover, CPD performance can be identified over the rCS range the practitioner might expect.
28. The implications of the difference between TAR and T_+ F-score will be addressed in future work.
29. All of the other measures are bound. In order to facilitate comparison, DOR was transformed from an "odds" format to the equivalent "probability" format.

References

[1] A. Jamain and D. J. Hand, "Mining supervised classification performance studies: a meta-analytic investigation," *Journal of Classification*, vol. 25, no. 1, pp. 87–112, 2008.

[2] R. P. W. Duin, "A note on comparing classifiers," *Pattern Recognition Letters*, vol. 17, no. 5, pp. 529–536, 1996.

[3] D. J. Hand, *Measurement Theory and Practice: The World Through Quantification*, Oxford University Press, New York, NY, USA, 2004.

[4] D. Böhning, W. Böhning, and H. Holling, "Revisiting Youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies," *Statistical Methods in Medical Research*, vol. 17, no. 6, pp. 543–554, 2008.

[5] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: an empirical analysis of supervised learning performance criteria," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, pp. 69–78, August 2004.

[6] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, pp. 233–240, June 2006.

[7] J. M. Fardy, "Evaluation of diagnostic tests," *Methods in Molecular Biology*, vol. 473, pp. 127–136, 2009.

[8] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, 2009.

[9] V. García, R. A. Mollineda, and J. S. Sánchez, "Theoretical analysis of a performance measure for imbalanced data," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*, pp. 617–620, Istanbul, Turkey, August 2010.

[10] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," *Communications in Computer and Information Science*, vol. 51, pp. 461–471, 2009.

[11] N. Japkowicz, "Why question machine learning evaluation methods?" in *Proceedings of the AAAI Evaluation Methods for Machine Learning Workshop*, pp. 6–11, July 2006.

[12] R. Potolea and C. Lemnaru, "A comprehensive study of the effect of class imbalance on the performance of classifiers," 2012, http://search.utcluj.ro/articole/Comprehensive Study.pdf.

[13] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *Proceedings of the AI 2006: Advances in Artificial Intelligence*, pp. 1015–1021, July 2006.

[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009.

[15] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[16] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of Clinical Epidemiology*, vol. 56, no. 11, pp. 1129–1135, 2003.

[17] D. D. Blakeley, E. Z. Oddone, V. Hasselblad, D. L. Simel, and D. B. Matchar, "Noninvasive carotid artery testing. A meta-analytic review," *Annals of Internal Medicine*, vol. 122, no. 5, pp. 360–367, 1995.

[18] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[19] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[20] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[21] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[22] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, vol. 232, no. 2, pp. 584–599, 1993.

[23] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*, pp. 3121–3124, Istanbul, Turkey, August 2010.

[24] A. Frank and A. Asuncion, "UCI machine learning repository," 2010, http://archive.ics.uci.edu/ml/.

[25] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[26] C. J. van Rijsbergen, "Information Retrieval," 1979, http://www.dcs.gla.ac.uk/Keith/Preface.html.

[27] C. W. Cleverdon, "The critical appraisal of information retrieval systems," 1968, http://hdl.handle.net/1826/1366.

[28] E. O. Cannon, A. Bender, D. S. Palmer, and J. B. O. Mitchell, "Chemoinformatics-based classification of prohibited substances employed for doping in sport," *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2369–2380, 2006.

[29] O. Carugo, "Detailed estimation of bioinformatics prediction reliability through the fragmented prediction performance plots," *BMC Bioinformatics*, vol. 8, article 380, 2007.

[30] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, and D. Plewczynski, "PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines," *Journal of Molecular Modeling*, vol. 17, no. 9, pp. 2191–2201, 2011.

[31] P. Dao, K. Wang, C. Collins, M. Ester, A. Lapuk, and S. C. Sahinalp, "Optimally discriminative subnetwork markers predict response to chemotherapy," *Bioinformatics*, vol. 27, no. 13, pp. i205–i213, 2011.

[32] K. K. Kandaswamy, K. C. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.

[33] T. Y. Lee, C. T. Lu, S. A. Chen et al., "Investigation and identification of protein-glutamyl carboxylation sites," in *Proceedings of the 10th International Conference on Bioinformatics. 1st ISCB Asia Joint Conference 2011: Bioinformatics*, 2011.

[34] G. Mirceva, A. Naumoski, and D. Davcev, "A novel fuzzy decision tree based method for detecting protein active sites," *Advances in Intelligent and Soft Computing*, vol. 150, pp. 51–60, 2012.

[35] M. S. Cline, K. Karplus, R. H. Lathrop, T. F. Smith, R. G. Rogers, and D. Haussler, "Information-theoretic dissection of pairwise contact potentials," *Proteins*, vol. 49, no. 1, pp. 7–14, 2002.

[36] C. Kauffman and G. Karypis, "An analysis of information content present in protein-DNA interactions," *Pacific Symposium on Biocomputing*, pp. 477–488, 2008.

[37] M. Kulharia, R. S. Goody, and R. M. Jackson, "Information theory-based scoring function for the structure-based prediction of protein-ligand binding affinity," *Journal of Chemical Information and Modeling*, vol. 48, no. 10, pp. 1990–1998, 2008.

[38] T. J. Magliery and L. Regan, "Sequence variation in ligand binding sites in proteins," *BMC Bioinformatics*, vol. 6, article 240, 2005.

[39] C. S. Miller and D. Eisenberg, "Using inferred residue contacts to distinguish between correct and incorrect protein models," *Bioinformatics*, vol. 24, no. 14, pp. 1575–1582, 2008.

[40] O. G. Othersen, A. G. Stefani, J. B. Huber, and H. Sticht, "Application of information theory to feature selection in protein docking," *Journal of Molecular Modeling*, vol. 18, no. 4, pp. 1285–1297, 2012.

[41] A. D. Solis and S. Rackovsky, "Information and discrimination in pairwise contact potentials," *Proteins*, vol. 71, no. 3, pp. 1071–1087, 2008.

[42] B. Sterner, R. Singh, and B. Berger, "Predicting and annotating catalytic residues: an information theoretic approach," *Journal of Computational Biology*, vol. 14, no. 8, pp. 1058–1073, 2007.

[43] A. M. Wassermann, B. Nisius, M. Vogt, and J. Bajorath, "Identification of descriptors capturing compound class-specific features by mutual information analysis," *Journal of Chemical Information and Modeling*, vol. 50, no. 11, pp. 1935–1940, 2010.

[44] J. Francois, H. Abdelnur, R. State, and O. Festor, "Ptf: passive temporal fingerprinting," in *Proceedings of the 12th IFIP/IEEE International Symposium on Integrated Network Management*, pp. 289–296, Dublin, UK, 2011.

[45] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, John Wiley & Sons, New York, NY, USA, 1991.

[46] R. W. Yeung, *A First Course in Information Theory. Information Technology: Transmission, Processing and Storage*, Kluwer Academic, New York, NY, USA, 2002.

[47] J. A. Swets, "Form of empirical ROCs in discrimination and diagnostic tasks. Implications for theory and measurement of performance," *Psychological Bulletin*, vol. 99, no. 2, pp. 181–198, 1986.

[48] J. A. Swets, "Indices of discrimination or diagnostic accuracy. Their ROCs and implied models," *Psychological Bulletin*, vol. 99, no. 1, pp. 100–117, 1986.

[49] D. Johnson, "Performance evaluation," 2003, http://cnx.org/content/m11274/1.3/content_info.

[50] J. M. Lobo, A. Jiménez-valverde, and R. Real, "AUC: a misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.

[51] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009.

[52] S. Vanderlooy and E. Hüllermeier, "A critical analysis of variants of the AUC," *Machine Learning*, vol. 72, no. 3, pp. 247–262, 2008.

[53] M. Majnik and Z. Bosnic, "ROC analysis of classifers in machine learning: survey," Tech. Rep. MM-1/2011, Faculty of Computer and Information Science, University of Ljubljana, 2011.

[54] K. Nishimura, D. Sugiyama, Y. Kogata et al., "Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis," *Annals of Internal Medicine*, vol. 146, no. 11, pp. 797–808, 2007.

[55] M. Schonlau, W. DuMouchel, W. H. Ju, A. F. Karr, M. Theus, and Y. Vardi, "Computer intrusion: detecting masquerades," *Statistical Science*, vol. 16, no. 1, pp. 58–74, 2001.

[56] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the 5th European Conference on Speech Communication and Technology*, pp. 1895–1898, Rhodes, Greece, 1997.

[57] I. K. Crombie and H. T. Davies, "What is meta-analysis?. 'What is ... ?'," series NPR09/1112, Hayward Medical Communications, 2009.

Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2013, Article ID 920157, 10 pages
http://dx.doi.org/10.1155/2013/920157

Research Article

Conservative Intensional Extension of Tarski's Semantics

Zoran Majkić

International Society for Research in Science and Technology, P.O. Box 2464, Tallahassee, FL 32316-2464, USA

Correspondence should be addressed to Zoran Majkić; majk.1234@yahoo.com

Received 30 May 2012; Revised 12 October 2012; Accepted 23 October 2012

Academic Editor: Konstantinos Lefkimmiatis

We considered an extension of the first-order logic (FOL) by Bealer's intensional abstraction operator. Contemporary use of the term "intension" derives from the traditional logical Frege-Russell doctrine that an idea (logic formula) has both an extension and an intension. Although there is divergence in formulation, it is accepted that the "extension" of an idea consists of the subjects to which the idea applies, and the "intension" consists of the attributes implied by the idea. From the Montague's point of view, the meaning of an idea can be considered as particular extensions in different possible worlds. In the case of standard FOL, we obtain a commutative homomorphic diagram, which is valid in each given possible world of an intensional FOL: from a free algebra of the FOL syntax, into its intensional algebra of concepts, and, successively, into an extensional relational algebra (different from Cylindric algebras). Then we show that this composition corresponds to the Tarski's interpretation of the standard extensional FOL in this possible world.

1. Introduction

In "Über Sinn und edeutung," Frege concentrated mostly on the senses of names, holding that all names have a sense (meaning). It is natural to hold that the same considerations apply to any expression that has an extension. But two general terms can have the same extension and different cognitive significance; two predicates can have the same extension and different cognitive significance; two sentences can have the same extension and different cognitive significance. So, general terms, predicates, and sentences all have senses as well as extensions. The same goes for any expression that has an extension or is a candidate for extension.

The significant aspect of an expression's meaning is its extension. We can stipulate that the extension of a sentence is its truth-value, and that the extension of a singular term is its referent. The extension of other expressions can be seen as associated entities that contribute to the truth-value of a sentence in a manner broadly analogous to the way in which the referent of a singular term contributes to the truth-value of a sentence. In many cases, the extension of an expression will be what we intuitively think of as its referent, although this need not hold in all cases. While Frege himself is often interpreted as holding that a sentence's referent is its truth-value, this claim is counterintuitive and widely disputed. We can avoid that issue in the present framework by using the technical term "extension." In this context, the claim that the extension of a sentence is its truth-value is a stipulation.

"Extensional" is most definitely a technical term. Say that the extension of a name is its denotation, the extension of a predicate is the set of things it applies to, and the extension of a sentence is its truth value. A logic is extensional if coextensional expressions can be substituted one for another in any sentence of the logic "salva veritate," that is, without a change in truth value. The intuitive idea behind this principle is that, in an extensional logic, the only logically significant notion of meaning that attaches to an expression is its extension. An intensional logics is exactly one in which substitutivity salva veritate fails for some of the sentences of the logic.

The first conception of intensional entities (or concepts) is built into the *possible-worlds* treatment of Properties, Relations, and Propositions (PRPs). This conception is commonly attributed to Leibniz and underlies Alonzo Church's alternative formulation of Frege's theory of senses ("*A formulation of the logic of sense and denotation*" in Henle, Kallen, and Langer, 3–24, and "*Outline of a revised formulation of the logic of sense and denotation*" in two parts, Nous, VII (1973), 24–33, and VIII, (1974), 135–156). This conception of PRPs is ideally suited for treating the *modalities* (necessity, possibility,

etc.) and to Montague's definition of intension of a given virtual predicate $\phi(x_1, \ldots, x_k)$ (a FOL open-sentence with the tuple of free variables $(x_1, \ldots x_k)$), as a mapping from possible worlds into extensions of this virtual predicate. Among the possible worlds, we distinguish the *actual* possible world. For example, if we consider a set of predicates, of a given Database, and their extensions in different time-instances, then the actual possible world is identified by the current instance of the time.

The second conception of intensional entities is to be found in Russell's doctrine of logical atomism. In this doctrine, it is required that all complete definitions of intensional entities be finite as well as unique and noncircular: it offers an *algebraic* way for definition of complex intensional entities from simple (atomic) entities (i.e., algebra of concepts), conception also evident in Leibniz's remarks. In a predicate logics, predicates and open-sentences (with free variables) express classes (properties and relations), and sentences express propositions. Note that classes (intensional entities) are *reified*, that is, they belong to the same domain as individual objects (particulars). This endows the intensional logics with a great deal of uniformity, making it possible to manipulate classes and individual objects in the same language. In particular, when viewed as an individual object, a class can be a member of another class.

The distinction between intensions and extensions is important (as in lexicography [1]), considering that extensions can be notoriously difficult to handle in an efficient manner. The extensional equality theory of predicates and functions under higher-order semantics (e.g., for two predicates with the same set of attributes, $p = q$ is true iff these symbols are interpreted by the same relation), that is, the strong equational theory of intensions, is not decidable, in general. For example, the second-order predicate calculus and Church's simple theory of types, both under the standard semantics, are not even semi-decidable. Thus, separating intensions from extensions makes it possible to have an equational theory over predicate and function names (intensions) that is separate from the extensional equality of relations and functions.

Relevant recent work about the intension, and its relationship with FOL, has been presented in [2] in the consideration of rigid and *nonrigid* objects, with respect to the possible worlds, where the rigid objects, like "George Washington," are the same things from possible world to possible world. Nonrigid objects, like "the Secretary-General of United Nations," are varying from circumstance to circumstance and can be modeled semantically by functions from possible worlds to domain of rigid objects, like intensional entities. But in his approach, differently from that one, fitting changes also the syntax of the FOL, by introducing an "extension of" operator, $\downarrow$, in order to distinguish the intensional entity "gross-domestic-product-of-Denmark," and its use in "the gross domestic product of Denmark is currently greater than gross domestic product of Finland." In his approach, if x is an intensional variable, $\downarrow x$ is extensional, while $\downarrow$ is not applicable to extensional variables, differently from our where each variable (concept) has both intensional and extension. Moreover, in his approach the problem arises because the action of letting x designate, that is, evaluating $\downarrow x$, and the action of passing to an alternative possible world, that is, of interpreting the existential modal operator $\Diamond$, are not actions that commute. To disambiguate this, one more piece of machinery is needed as well, which substantially and ad-hock changes the syntax and semantics of FOL, introduces the Higher-order Modal logics, and is not a conservative extension of Tarski's semantics.

In most recent work in [3, 4] it is given an intensional version of first-order *hybrid* logic, which is also a hybridized version of Fitting's intensional FOL, by a kind of generalized models, thus, different from our approaches to conservative extension of Tarski's semantics to intensional FOL.

Another recent relevant work is presented by I-logic in [5], which combines both approach to semantics of intensional objects of Montague and Fitting.

We recall that Intensional Logic Programming is a new form of logic programming based on intensional logic and possible worlds semantics and is a well-defined practice in using the intensional semantics [6]. Intensional logic allows us to use logic programming to specify nonterminating computations and to capture the dynamic aspects of certain problems in a natural and problem-oriented style. The meanings of formulas of an intensional first-order language are given according to intensional interpretations and to elements of a set of possible worlds. Neighborhood semantics is employed as an abstract formulation of the denotations of intensional operators. The model-theoretic and fixpoint semantics of intensional logic programs are developed in terms of least (minimum) intensional Herbrand models. Intensional logic programs with intensional operator definitions are regarded as metatheories.

In what follows, we denote by B^A the set of all functions from A to B, and by A^n an n-folded cartesian product $A \times \cdots \times A$ for $n \geq 1$. By f, t we denote empty set $\emptyset$ and singleton set $\{\langle\,\rangle\}$, respectively (with the empty tuple $\langle\,\rangle$ i.e., the unique tuple of 0-ary relation), which may be thought of as falsity f and truth t, as those used in the relational algebra. For a given domain $\mathscr{D}$, we define that $\mathscr{D}^0$ is a singleton set $\{\langle\,\rangle\}$, so that $\{f, t\} = \mathscr{P}(\mathscr{D}^0)$, where $\mathscr{P}$ is the powerset operator.

2. Intensional FOL Language with Intensional Abstraction

Intensional entities are such concepts as propositions and properties. The term "intensional" means that they violate the principle of extensionality, the principle that extensional equivalence implies identity. All (or most) of these intensional entities have been classified at one time or another as kinds of Universals [7].

We consider a nonempty domain $\mathscr{D} = D_{-1} \bigcup D_I$, where a subdomain D_{-1} is made of particulars (extensional entities), and the rest $D_I = D_0 \bigcup D_1 \cdots \bigcup D_n \cdots$ is made of universals (D_0 for propositions (the 0-ary concepts)), and D_n, $n \geq 1$, for n-ary concepts.

The fundamental entities are *intensional abstracts* or so-called, that-clauses. We assume that they are singular terms; Intensional expressions like "believe," "mean," "assert,"

"know," are standard two-place predicates that take "that"-clauses as arguments. Expressions like "is necessary," "is true," and "is possible" are one-place predicates that take "that"-clauses as arguments. For example, in the intensional sentence "it is necessary that ϕ," where ϕ is a proposition, the "that ϕ" is denoted by the $\prec \phi \succ$, where $\prec\succ$ is the intensional abstraction operator, which transforms a logic formula into a *term*. Or, for example, "x believes that ϕ" is given by formula $p_i^2(x, \prec \phi \succ)$ (p_i^2 is binary "believe" predicate).

Here we will present an intensional FOL with slightly different intensional abstraction than that originally presented in [8].

Definition 1. The syntax of the first-order logic language with intensional abstraction $\prec\succ$, denoted by $\mathscr{L}$, is as follows:

logic operators ($\wedge, \neg, \exists$), predicate letters in P (functional letters is considered as particular case of predicate letters), variables $x, y, z, \ldots$ in $\mathscr{V}$, abstraction $\prec \_ \succ$, and punctuation symbols (comma, parenthesis). With the following simultaneous inductive definition of *term* and *formula*,

(1) all variables and constants (0-ary functional letters in P) are terms;

(2) if $t_1, \ldots, t_k$ are terms, then $p_i^k(t_1, \ldots, t_k)$ is a formula ($p_i^k \in P$ is a k-ary predicate letter);

(3) if ϕ and ψ are formulae, then $(\phi \wedge \psi)$, $\neg\phi$, and $(\exists x)\phi$ are formulae;

(4) if $\phi(\mathbf{x})$ is a formula (virtual predicate) with a list of free variables in $\mathbf{x} = (x_1, \ldots, x_n)$ (with ordering from-left-to-right of their appearance in ϕ), and α is its sublist of *distinct* variables, then $\prec \phi \succ_\alpha^\beta$ is a term, where β is the remaining list of free variables preserving ordering in $\mathbf{x}$ as well. The externally quantifiable variables are the *free* variables not in α. When $n = 0, \prec \phi \succ$ is a term that denotes a proposition, for $n \geq 1$ it denotes an n-ary concept.

An occurrence of a variable x_i in a formula (or a term) is *bound* (*free*) if and only if it lies (does not lie) within a formula of the form $(\exists x_i)\phi$ (or a term of the form $\prec \phi \succ_\alpha^\beta$ with $x_i \in \alpha$). A variable is free (bound) in a formula (or term) if and only if it has (does not have) a free occurrence in that formula (or term).

A *sentence* is a formula having no free variables. The binary predicate letter p_1^2 for identity is singled out as a distinguished logical predicate, and formulae of the form $p_1^2(t_1, t_2)$ are to be rewritten in the form $t_1 \doteq t_2$. We denote by $R_=$ the binary relation obtained by standard Tarski's interpretation of this predicate p_1^2. The logic operators $\forall, \vee, \Rightarrow$ are defined in terms of $(\wedge, \neg, \exists)$ in the usual way.

Remark 2. The k-ary functional symbols, for $k \geq 1$, in standard (extensional) FOL are considered as $(k+1)$-ary predicate symbols p^{k+1}: the function $f : \mathscr{D}^k \to \mathscr{D}$ is considered as a relation obtained from its graph $R = \{(d_1, \ldots, d_k, f(d_1, \ldots, d_k)) \mid d_i \in \mathscr{D}\}$, represented by a predicate symbol p^{k+1}.

The universal quantifier is defined by $\forall = \neg\exists\neg$. Disjunction and implication are expressed by $\phi \vee \psi = \neg(\neg\phi \wedge \neg\psi)$ and $\phi \Rightarrow \psi = \neg\phi \vee \psi$. In FOL with the identity $\doteq$, the formula $(\exists_1 x)\phi(x)$ denotes the formula $(\exists x)\phi(x) \wedge (\forall x)(\forall y)(\phi(x) \wedge \phi(y) \Rightarrow (x \doteq y))$. We denote by $R_=$ the Tarski's interpretation of $\doteq$.

In what follows, any open-sentence, a formula ϕ with nonempty tuple of free variables $(x_1, \ldots, x_m)$, will be called a m-ary *virtual predicate*, denoted also by $\phi(x_1, \ldots, x_m)$. This definition contains the precise method of establishing the *ordering* of variables in this tuple: such a method that will be adopted here is the ordering of appearance, from left to right, of free variables in ϕ. This method of composing the tuple of free variables is the unique and canonical way of definition of the virtual predicate from a given formula.

An *intensional interpretation* of this intensional FOL is a mapping between the set $\mathscr{L}$ of formulae of the logic language and intensional entities in $\mathscr{D}$, $I : \mathscr{L} \to \mathscr{D}$, which is a kind of "conceptualization", such that an open-sentence (virtual predicate) $\phi(x_1, \ldots, x_k)$ with a tuple of all free variables $(x_1, \ldots, x_k)$ is mapped into a k-ary *concept*, that is, an intensional entity $u = I(\phi(x_1, \ldots, x_k)) \in D_k$, and (closed) sentence ψ into a proposition (i.e., *logic* concept) $v = I(\psi) \in D_0$ with $I(\top) = Truth \in D_0$ for a FOL tautology $\top$. A language constant c is mapped into a particular (an extensional entity) $a = I(c) \in D_{-1}$ if it is a proper name, otherwise in a correspondent concept in $\mathscr{D}$.

An assignment $g : \mathscr{V} \to \mathscr{D}$ for variables in $\mathscr{V}$ is applied only to free variables in terms and formulae. Such an assignment $g \in \mathscr{D}^{\mathscr{V}}$ can be recursively uniquely extended into the assignment $g^* : \mathscr{T} \to \mathscr{D}$, where $\mathscr{T}$ denotes the set of all terms (here I is an intensional interpretation of this FOL, as explained in what follows), by

(1) $g^*(t) = g(x) \in \mathscr{D}$ if the term t is a variable $x \in \mathscr{V}$;

(2) $g^*(t) = I(c) \in \mathscr{D}$ if the term t is a constant $c \in P$;

(3) if t is an abstracted term $\prec \phi \succ_\alpha^\beta$, then $g^*(\prec \phi \succ_\alpha^\beta) = I(\phi[\beta/g(\beta)]) \in D_k, k = |\alpha|$ (i.e., the number of variables in α), where $g(\beta) = g(y_1, \ldots, y_m) = (g(y_1), \ldots, g(y_m))$ and $[\beta/g(\beta)]$ is a uniform replacement of each ith variable in the list β with the ith constant in the list $g(\beta)$. Notice that α is the list of all free variables in the formula $\phi[\beta/g(\beta)]$.

We denote by t/g (or ϕ/g) the ground term (or formula) without free variables, obtained by assignment g from a term t (or a formula ϕ), and by $\phi[x/t]$ the formula obtained by uniformly replacing x by a term t in ϕ.

The distinction between intensions and extensions is important especially because we are now able to have and *equational theory* over intensional entities (as $\prec \phi \succ$), that is, predicate and function "names," which is separate from the extensional equality of relations and functions. An *extensionalization function* h assigns to the intensional elements of $\mathscr{D}$ an appropriate extension as follows: for each proposition $u \in D_0$, $h(u) \in \{f, t\} \subseteq \mathscr{P}(D_{-1})$ is its extension (true or false value); for each n-ary concept $u \in D_n$, $h(u)$ is a subset of $\mathscr{D}^n$ (nth Cartesian product of $\mathscr{D}$); in the case of particulars $u \in D_{-1}$, $h(u) = u$.

The sets f, t are empty set $\{\}$ and set $\{\langle\rangle\}$ (with the empty tuple $\langle\rangle \in D_{-1}$, i.e., the unique tuple of 0-ary relation)

which may be thought of as falsity and truth, as those used in the Codd's relational-database algebra [9], respectively, while $Truth \in D_0$ is the concept (intension) of the tautology.

We define that $\mathscr{D}^0 = \{\langle\,\rangle\}$, so that $\{f, t\} = \mathscr{P}(\mathscr{D}^0)$, where $\mathscr{P}$ is the powerset operator. Thus we have (we denote the disjoint union by "+"):

$$h = \left(h_{-1} + \sum_{i\geq 0} h_i\right) : \sum_{i\geq -1} D_i \longrightarrow D_{-1} + \sum_{i\geq 0} \mathscr{P}\left(D^i\right), \quad (1)$$

where $h_{-1} = id : D_{-1} \to D_{-1}$ is identity mapping, the mapping $h_0 : D_0 \to \{f, t\}$ assigns the truth values in $\{f, t\}$ to all propositions, and the mappings $h_i : D_i \to \mathscr{P}(D^i)$, $i \geq 1$, assign an extension to all concepts. Thus, the intensions can be seen as *names* of abstract or concrete entities, while the extensions correspond to various rules that these entities play in different worlds.

Remark 3 (Tarski's constraints). This intensional semantics has to preserve standard Tarski's semantics of the FOL. That is, for any formula $\phi \in \mathscr{L}$ with a tuple of free variables $(x_1, \ldots, x_k)$, and $h \in \mathscr{E}$, the following conservative conditions for all assignments $g, g\prime \in \mathscr{D}^{\mathscr{V}}$ have to be satisfied:

(T) $h(I(\phi/g)) = t$ if and only if $(g(x_1), \ldots, g(x_k)) \in h(I(\phi))$ and if ϕ is a predicate letter p^k, $k \geq 2$ which represents a $(k-1)$-ary functional symbol f^{k-1} in standard FOL,

(TF) $h(I(\phi/g)) = h(I(\phi/g')) = t$, and $\forall_{1\leq i\leq k-1}(g'(x_i) = g(x_i))$ implies $g'(x_{k+1}) = g(x_{k+1})$.

Thus, intensional FOL has a simple Tarski first-order semantics, with a decidable unification problem, but we need also the actual world mapping which maps any intensional entity to its *actual world extension*. In what follows, we will identify a *possible world* by a particular mapping which assigns, in such a possible world, the extensions to intensional entities. This is a direct bridge between an intensional FOL and a possible worlds representation [10–15], where the intension (meaning) of a proposition is a *function*, from a set of possible $\mathscr{W}$ worlds into the set of truth values. Consequently, $\mathscr{E}$ denotes the set of possible *extensionalization functions* h satisfying the constraint (T). Each $h \in \mathscr{E}$ may be seen as a *possible world* (analogously to Montague's intensional semantics for natural language [12, 14]), as it has been demonstrated in [16, 17] and given by the bijection $is : \mathscr{W} \simeq \mathscr{E}$.

Now we are able to define formally this intensional semantics [15].

Definition 4. A two-step intensional semantics.

Let $\mathfrak{R} = \bigcup_{k\in\mathbb{N}} \mathscr{P}(\mathscr{D}^k) = \sum_{k\in\mathbb{N}} \mathscr{P}(D^k)$ be the set of all k-ary relations, where $k \in \mathbb{N} = \{0, 1, 2, \ldots\}$. Notice that $\{f, t\} = \mathscr{P}(\mathscr{D}^0) \in \mathfrak{R}$, that is, the truth values are extensions in $\mathfrak{R}$.

The intensional semantics of the logic language with the set of formulae $\mathscr{L}$ can be represented by the mapping

$$\mathscr{L} \xrightarrow{I} \mathscr{D} \Longrightarrow_{w\in\mathscr{W}} \mathfrak{R}, \quad (2)$$

where $\xrightarrow{I}$ is a *fixed intensional* interpretation $I : \mathscr{L} \to \mathscr{D}$ and $\Rightarrow_{w\in\mathscr{W}}$ is *the set* of all extensionalization functions $h = is(w) : \mathscr{D} \to \mathfrak{R}$ in $\mathscr{E}$, where $is : \mathscr{W} \to \mathscr{E}$ is the mapping from the set of possible worlds to the set of extensionalization functions.

We define the mapping $I_n : \mathscr{L}_{op} \to \mathfrak{R}^{\mathscr{W}}$, where $\mathscr{L}_{op}$ is the subset of formulae with free variables (virtual predicates), such that for any virtual predicate $\phi(x_1, \ldots, x_k) \in \mathscr{L}_{op}$ the mapping $I_n(\phi(x_1, \ldots, x_k)) : \mathscr{W} \to \mathfrak{R}$ is the Montague's meaning (i.e., *intension*) of this virtual predicate [10–14], that is, the mapping which returns with the extension of this (virtual) predicate in each possible world $w \in \mathscr{W}$.

We adopted this two-step intensional semantics, instead of well-known Montague's semantics (which lies in the construction of a compositional and recursive semantics that covers both intension and extension), because of a number of weakness of the second semantics:

Example 5. Let us consider the following two past participles: "bought" and "sold" (with unary predicates $p_1^1(x)$, "x has been bought", and $p_2^1(x)$, "x has been sold"). These two different concepts in the Montague's semantics would have not only the same extension but also their intension, from the fact that their extensions are identical in every possible world.

Within the two-step formalism, we can avoid this problem by assigning two different concepts (meanings) $u = I(p_1^1(x))$ and $v = I(p_2^1(x))$ in $\in D_1$. Note that we have the same problem in the Montague's semantics for two sentences with different meanings, which bear the same truth value across all possible worlds: in Montague's semantics, they will be forced to the *same* meaning.

Another relevant question with respect to this two-step interpretations of an intensional semantics is how in it the extensional identity relation $\doteq$ (binary predicate of the identity) of the FOL is managed. Here this extensional identity relation is mapped into the binary concept $Id = I(\doteq (x, y)) \in D_2$, such that $(\forall w \in \mathscr{W})(is(w)(Id) = R_=)$, where $\doteq (x, y)$ (i.e., $p_1^2(x, y)$) denotes an atom of the FOL of the binary predicate for identity in FOL, usually written by FOL formula $x \doteq y$.

Note that here we prefer to distinguish this *formal symbol* $\doteq \in P$ of the built-in identity binary predicate letter in the FOL, from the standard mathematical symbol "=" used in all mathematical definitions in this paper.

In what follows, we will use the function $f_{\langle\rangle} : \mathfrak{R} \to \mathfrak{R}$, such that for any relation $R \in \mathfrak{R}$, $f_{\langle\rangle}(R) = \{\langle\,\rangle\}$ if $R \neq \emptyset$; $\emptyset$ otherwise. Let us define the following set of algebraic operators for relations in $\mathfrak{R}$.

(1) Binary operator $\bowtie_S : \mathfrak{R} \times \mathfrak{R} \to \mathfrak{R}$, such that for any two relations $R_1, R_2 \in \mathfrak{R}$, the $R_1 \bowtie_S R_2$ is equal to the relation obtained by natural join of these two relations if S is a nonempty set of pairs of joined columns of respective relations (where the first argument is the column index of the relation R_1 while the second argument is the column index of the joined column of

the relation R_2); otherwise it is equal to the cartesian product $R_1 \times R_2$.

For example, the logic formula $\phi(x_i, x_j, x_k, x_l, x_m) \wedge \psi(x_l, y_i, x_j, y_j)$ will be traduced by the algebraic expression $R_1 \bowtie_S R_2$ where $R_1 \in \mathscr{P}(\mathscr{D}^5)$, $R_2 \in \mathscr{P}(\mathscr{D}^4)$ are the extensions for a given Tarski's interpretation of the virtual predicate ϕ, ψ relatively, so that $S = \{(4,1),(2,3)\}$ and the resulting relation will have the following ordering of attributes: $(x_i, x_j, x_k, x_l, x_m, y_i, y_j)$.

(2) Unary operator $\sim$: $\mathfrak{R} \to \mathfrak{R}$, such that for any k-ary (with $k \geq 0$) relation $R \in \mathscr{P}(\mathscr{D}^k) \subset \mathfrak{R}$, we have that $\sim(R) = \mathscr{D}^k \setminus R \in \mathscr{D}^k$, where "\" is the substraction of relations. For example, the logic formula $\neg\phi(x_i, x_j, x_k, x_l, x_m)$ will be traduced by the algebraic expression $\mathscr{D}^5 \setminus R$ where R is the extensions for a given Tarski's interpretation of the virtual predicate ϕ.

(3) Unary operator $\pi_{-m} : \mathfrak{R} \to \mathfrak{R}$, such that for any k-ary (with $k \geq 0$) relation $R \in \mathscr{P}(\mathscr{D}^k) \subset \mathfrak{R}$, we have that $\pi_{-m}(R)$ is equal to the relation obtained by elimination of the mth column of the relation R if $1 \leq m \leq k$ and $k \geq 2$; equal to $f_{\langle\rangle}(R)$ if $m = k = 1$; otherwise it is equal to R.

For example, the logic formula $(\exists x_k)\phi(x_i, x_j, x_k, x_l, x_m)$ will be traduced by the algebraic expression $\pi_{-3}(R)$ where R is the extensions for a given Tarski's interpretation of the virtual predicate ϕ and the resulting relation will have the following ordering of attributes: (x_i, x_j, x_l, x_m).

Notice that the ordering of attributes of resulting relations corresponds to the method used for generating the ordering of variables in the tuples of free variables adopted for virtual predicates.

Analogously to Boolean algebras, which are extensional models of propositional logic, we introduce now an intensional algebra for this intensional FOL, as follows.

Definition 6. Intensional algebra for the intensional FOL in Definition 1 is a structure $\mathscr{A}_{\text{int}} = (\mathscr{D}, f, t, Id, Truth, \{\text{con}j_S\}_{S \in \mathscr{P}(\mathbb{N}^2)}, \text{neg}, \{\text{exists}_n\}_{n \in \mathbb{N}})$, with binary operations $\text{con}j_S : D_I \times D_I \to D_I$, unary operation $\text{neg} : D_I \to D_I$, unary operations $\text{exists}_n : D_I \to D_I$, such that for any extensionalization function $h \in \mathscr{E}$, and $u \in D_k, v \in D_j, k, j \geq 0$,

(1) $h(Id) = R_=$ and $h(Truth) = \{\langle\rangle\}$.

(2) $h(\text{con}j_S(u,v)) = h(u) \bowtie_S h(v)$, where $\bowtie_S$ is the natural join operation defined above and $\text{con}j_S(u,v) \in D_m$ where $m = k + j - |S|$ if for every pair $(i_1, i_2) \in S$ it holds that $1 \leq i_1 \leq k$, $1 \leq i_2 \leq j$ (otherwise $\text{con}j_S(u,v) \in D_{k+j}$).

(3) $h(\text{neg}(u)) = \sim(h(u)) = \mathscr{D}^k \setminus (h(u))$, where $\sim$ is the operation defined above and $\text{neg}(u) \in D_k$.

(4) $h(\text{exists}_n(u)) = \pi_{-n}(h(u))$, where π_{-n} is the operation defined above and $\text{exists}_n(u) \in D_{k-1}$ if $1 \leq n \leq k$ (otherwise exists_n is the identity function).

Notice that for $u \in D_0$, $h(\text{neg}(u)) = \sim(h(u)) = \mathscr{D}^0 \setminus (h(u)) = \{\langle\rangle\} \setminus (h(u)) \in \{f, t\}$.

We define a derived operation union : $(\mathscr{P}(D_i) \setminus \emptyset) \to D_i$, $i \geq 0$, such that, for any $B = \{u_1, \ldots, u_n\} \in \mathscr{P}(D_i)$ we have that $\text{union}(\{u_1, \ldots, u_n\}) =_{\text{def}} u_1$ if $n = 1$; $\text{neg}(\text{con}j_S(\text{neg}(u_1), \text{con}j_S(\ldots, \text{neg}(u_n))\ldots)$, where $S = \{(l,l) | 1 \leq l \leq i\}$, otherwise. Than we obtain that for $n \geq 2$,

$$\begin{aligned} & h(\text{union}(B)) \\ &= h\left(\text{neg}\left(\text{con}j_S\left(\text{neg}\left(u_1\right), \text{con}j_S\left(\ldots, \text{neg}\left(u_n\right)\right)\right)\ldots\right)\right. \\ &= \mathscr{D}^i \setminus \left(\left(\mathscr{D}^i \setminus h\left(u_1\right)\right) \bowtie_S \cdots \bowtie_S \left(\mathscr{D}^i \setminus h\left(u_n\right)\right)\right) \\ &= \mathscr{D}^i \setminus \left(\left(\mathscr{D}^i \setminus h\left(u_1\right)\right) \bigcap \cdots \bigcap \left(\mathscr{D}^i \setminus h\left(u_n\right)\right)\right) \\ &= \bigcup \left\{h\left(u_j\right) \mid 1 \leq j \leq n\right\} = \bigcup \{h(u) \mid u \in B\}. \end{aligned} \tag{3}$$

Intensional interpretation $I : \mathscr{L} \to \mathscr{D}$ satisfies the following homomorphic extension.

(1) The logic formula $\phi(x_i, x_j, x_k, x_l, x_m) \wedge \psi(x_l, y_i, x_j, y_j)$ will be intensionally interpreted by the concept $u_1 \in D_7$, obtained by the algebraic expression $\text{con}j_S(u,v)$ where $u = I(\phi(x_i, x_j, x_k, x_l, x_m)) \in D_5, v = I(\psi(x_l, y_i, x_j, y_j)) \in D_4$ are the concepts of the virtual predicates ϕ, ψ, relatively, and $S = \{(4,1),(2,3)\}$. Consequently, we have that for any two formulae $\phi, \psi \in \mathscr{L}$ and a particular operator $\text{con}j_S$ uniquely determined by tuples of free variables in these two formulae, $I(\phi \wedge \psi) = \text{con}j_S(I(\phi), I(\psi))$.

(2) The logic formula $\neg\phi(x_i, x_j, x_k, x_l, x_m)$ will be intensionally interpreted by the concept $u_1 \in D_5$, obtained by the algebraic expression $\text{neg}(u)$ where $u = I(\phi(x_i, x_j, x_k, x_l, x_m)) \in D_5$ is the concept of the virtual predicate ϕ. Consequently, we have that for any formula $\phi \in \mathscr{L}$, $I(\neg\phi) = \text{neg}(I(\phi))$.

(3) The logic formula $(\exists x_k)\phi(x_i, x_j, x_k, x_l, x_m)$ will be intensionally interpreted by the concept $u_1 \in D_4$, obtained by the algebraic expression $\text{exists}_3(u)$ where $u = I(\phi(x_i, x_j, x_k, x_l, x_m)) \in D_5$ is the concept of the virtual predicate ϕ. Consequently, we have that for any formula $\phi \in \mathscr{L}$ and a particular operator exists_n uniquely determined by the position of the existentially quantified variable in the tuple of free variables in ϕ (otherwise $n = 0$ if this quantified variable is not a free variable in ϕ), $I((\exists x)\phi) = \text{exists}_n(I(\phi))$.

Once one has found a method for specifying the interpretations of singular terms of $\mathscr{L}$ (take in consideration the particularity of abstracted terms), the Tarski-style definitions of truth and validity for $\mathscr{L}$ may be given in the customary way. What is proposed specifically is a method for characterizing the intensional interpretations of singular terms of $\mathscr{L}$ in such a way that a given singular abstracted term $\prec \phi \succ_\alpha^\beta$ will denote an appropriate property, relation, or proposition, depending on the value of $m = |\alpha|$. Thus, the mapping of intensional abstracts (terms) into $\mathscr{D}$ will be defined differently from that given in the version of Bealer [18], as follows.

Definition 7. An intensional interpretation I can be extended to abstracted terms as follows: for any abstracted term $\prec \phi \succ_{\alpha}^{\beta}$, we define that

$$I\left(\prec \phi \succ_{\alpha}^{\beta}\right) = \text{union}\left(\left\{I\left(\phi\left[\frac{\beta}{g(\beta)}\right]\right) \mid g \in \mathscr{D}^{\overline{\beta}}\right\}\right), \quad (4)$$

where $\overline{\beta}$ denotes the set of elements in the list β, and the assignments in $\mathscr{D}^{\overline{\beta}}$ are limited only to the variables in $\overline{\beta}$.

Remark 8. Here we can make the question if there is a sense to extend the interpretation also to (abstracted) terms, because in Tarski's interpretation of FOL we do not have any interpretation for terms, but only the assignments for terms, as we defined previously by the mapping $g^{*}: \mathscr{T} \to \mathscr{D}$. The answer is positive, because the abstraction symbol $\prec \_ \succ_{\alpha}^{\beta}$ can be considered as a kind of the unary built-in functional symbol of intensional FOL, so that we can apply the Tarski's interpretation to this functional symbol into the fixed mapping $I(\prec \_ \succ_{\alpha}^{\beta}): \mathscr{L} \to \mathscr{D}$, so that for any $\phi \in \mathscr{L}$ we have that $I(\prec \phi \succ_{\alpha}^{\beta})$ is equal to the application of this function to the value ϕ, that is, to $I(\prec \_ \succ_{\alpha}^{\beta})(\phi)$. In such an approach, we would introduce also the typed variable X for the formulae in $\mathscr{L}$, so that the Tarski's assignment for this functional symbol with variable X, with $g(X) = \phi \in \mathscr{L}$, can be given by

$$\begin{aligned} & g^{*}\left(\prec \_ \succ_{\alpha}^{\beta}(X)\right) \\ & = I\left(\prec \_ \succ_{-\alpha}^{\beta}\right)(g(X)) = I\left(\prec \_ \succ_{\alpha}^{\beta}\right)(\phi) \\ & = \begin{cases} \langle\rangle \in D_{-1}, \\ \text{if } \overline{\alpha} \bigcup \overline{\beta} \text{ is not equal to the set of free variables in } \phi; \\ = \text{union}\left(\left\{I\left(\phi\left[\dfrac{\beta}{g'(\beta)}\right]\right) \mid g' \in \mathscr{D}^{\overline{\beta}}\right\}\right) \in D_{|\overline{\alpha}|}, \\ \qquad\qquad \text{otherwise.} \end{cases} \end{aligned} \quad (5)$$

Notice than if $\beta = \emptyset$ is the empty list, then $I(\prec \phi \succ_{\alpha}^{\beta}) = I(\phi)$. Consequently, the denotation of $\prec \phi \succ$ is equal to the meaning of a proposition ϕ, that is, $I(\prec \phi \succ) = I(\phi) \in D_0$. In the case when ϕ is an atom $p_i^m(x_1, \ldots, x_m)$, then $I(\prec p_i^m(x_1, \ldots, x_m) \succ_{x_1, \ldots, x_m}) = I(p_i^m(x_1, \ldots, x_m)) \in D_m$, while $I(\prec p_i^m(x_1, \ldots, x_m) \succ^{x_1, \ldots, x_m}) = \text{union}(\{I(p_i^m(g(x_1), \ldots, g(x_m))) \mid g \in \mathscr{D}^{\{x_1, \ldots, x_m\}}\}) \in D_0$, with $h(I(\prec p_i^m(x_1, \ldots, x_m) \succ^{x_1, \ldots, x_m})) = h(I((\exists x_1) \cdots (\exists x_m)\, p_i^m(x_1, \ldots, x_m))) \in \{f, t\}$.

For example,

$$\begin{aligned} & h\left(I\left(\prec p_i^1(x_1) \wedge \neg p_i^1(x_1) \succ^{x_1}\right)\right) \\ & \quad = h\left(I\left((\exists x_1)\left(\prec p_i^1(x_1) \wedge \neg p_i^1(x_1) \succ^{x_1}\right)\right)\right) = f. \end{aligned} \quad (6)$$

The interpretation of a more complex abstract $\prec \phi \succ_{\alpha}^{\beta}$ is defined in terms of the interpretations of the relevant syntactically simpler expressions, because the interpretation of more complex formulae is defined in terms of the interpretation of the relevant syntactically simpler formulae, based on the intensional algebra above. For example, $I(p_i^1(x) \wedge p_k^1(x)) = \text{con}j_{\{(1),1)\}}(I(p_i^1(x)), I(p_k^1(x)))$, $I(\neg\phi) = \text{neg}(I(\phi))$, $I(\exists x_i)\phi(x_i, x_j, x_i, x_k) = \text{exists}_3(I(\phi))$.

Consequently, based on the intensional algebra in Definition 6 and on intensional interpretations of abstracted terms in Definition 7, it holds that the interpretation of any formula in $\mathscr{L}$ (and any abstracted term) will be reduced to an algebraic expression over interpretations of primitive atoms in $\mathscr{L}$. This obtained expression is finite for any finite formula (or abstracted term) and represents the *meaning* of such finite formula (or abstracted term).

The *extension* of an abstracted term satisfy the following property.

Proposition 9. *For any abstracted term $\prec \phi \succ_{\alpha}^{\beta}$ with $|\alpha| \geq 1$, we have that*

$$h\left(I\left(\prec \phi \succ_{\alpha}^{\beta}\right)\right) = \pi_{-\beta}\left(h\left(I\left(\phi\right)\right)\right), \quad (7)$$

where $\pi_{-(y_1, \ldots, y_k)} = \pi_{-y_1} \circ \cdots \circ \pi_{-y_1}$, $\circ$ is the sequential composition of functions, and $\pi_{-\emptyset}$ is an identity.

Proof. Let $\mathbf{x}$ be a tuple of all free variables in ϕ, so that $\overline{\mathbf{x}} = \overline{\alpha} \bigcup \overline{\beta}$, $\alpha = (x_1, \ldots, x_k)$, then we have that $h(I(\prec \phi \succ_{\alpha}^{\beta})) = h(\text{union}(\{I(\phi[\beta/g(\beta)]) \mid g \in \mathscr{D}^{\overline{\beta}}\}))$, from Definition 7 $= \bigcup\{h(I(\phi[\beta/g(\beta)])) \mid g \in \mathscr{D}^{\overline{\beta}}\} = \bigcup\{\{(g_1(x_1), \ldots, g_1(x_k)) \mid g_1 \in \mathscr{D}^{\overline{\alpha}} \text{ and } h(I(\phi[\beta/g(\beta)][\alpha/g_1(\alpha)])) = t\} \mid g \in \mathscr{D}^{\overline{\beta}}\} = \{g_1(\alpha) \mid g_1 \in \mathscr{D}^{\overline{\alpha} \cup \overline{\beta}} \text{ and } h(I(\phi/g_1)) = t\} = \pi_{-\beta}(\{g_1(\mathbf{x}) \mid g_1 \in \mathscr{D}^{\overline{\mathbf{x}}} \text{ and } h(I(\phi/g_1)) = t\}) = \pi_{-\beta}(\{g_1(\mathbf{x}) \mid g_1 \in \mathscr{D}^{\overline{\mathbf{x}}} \text{ and } g_1(\mathbf{x}) \in h(I(\phi))\})$, by $(\mathbf{T}) = \pi_{-\beta}(h(I(\phi)))$. □

We can correlate $\mathscr{E}$ with a possible-world semantics. Such a correspondence is a natural identification of intensional logics with modal Kripke-based logics.

Definition 10 (model). A model for intensional FOL with fixed intensional interpretation I, which expresses the two-step intensional semantics in Definition 4, is the Kripke structure $\mathscr{M}_{\text{int}} = (\mathscr{W}, \mathscr{D}, V)$, where $\mathscr{W} = \{is^{-1}(h) | h \in \mathscr{E}\}$, a mapping $V: \mathscr{W} \times P \to \bigcup_{n<\omega}\{t, f\}^{\mathscr{D}^n}$, with P a set of predicate symbols of the language, such that for any world $w = is^{-1}(h) \in \mathscr{W}$, $p_i^n \in P$, and $(u_1, \ldots, u_n) \in \mathscr{D}^n$ it holds that $V(w, p_i^n)(u_1, \ldots, u_n) = h(I(p_i^n(u_1, \ldots, u_n)))$. The satisfaction relation $\vDash_{w,g}$ for a given $w \in \mathscr{W}$ and assignment $g \in \mathscr{D}^{\mathscr{V}}$ is defined as follows:

(1) $\mathscr{M} \vDash_{w,g} p_i^k(x_1, \ldots, x_k)$ if and only if $V(w, p_i^k)(g(x_1), \ldots, g(x_k)) = t$,

(2) $\mathscr{M} \vDash_{w,g} \varphi \wedge \phi$ if and only if $\mathscr{M} \vDash_{w,g} \varphi$ and $\mathscr{M} \vDash_{w,g} \phi$,

(3) $\mathscr{M} \vDash_{w,g} \neg\varphi$ if and only if not $\mathscr{M} \vDash_{w,g} \varphi$,

(4) $\mathscr{M} \vDash_{w,g} (\exists x)\phi$ if and only if

(4.1) $\mathscr{M} \vDash_{w,g} \phi$, if x is not a free variable in ϕ;

(4.2) exists $u \in \mathscr{D}$ such that $\mathscr{M} \vDash_{w,g} \phi[x/u]$, if x is a free variable in ϕ.

It is easy to show that the satisfaction relation $\models$ for this Kripke semantics in a world $w = is^{-1}(h)$ is defined by $\mathcal{M} \models_{w,g} \phi$ if and only if $h(I(\phi/g)) = t$.

We can enrich this intensional FOL by another modal operators, as, for example, the "necessity" universal logic operator $\Box$ with accessibility relation $\mathcal{R} = \mathcal{W} \times \mathcal{W}$, obtaining an S5 Kripke structure $\mathcal{M}_{\text{int}} = (\mathcal{W}, \mathcal{R}, \mathcal{D}, V)$. In this case, we are able to define the following equivalences between the abstracted terms without free variables $< \phi >_{\alpha}^{\beta_1}/g$ and $< \psi >_{\alpha}^{\beta_2}/g$, where all free variables (not in α) are instantiated by $g \in \mathcal{D}^{\mathcal{V}}$ (here $A \equiv B$ denotes the formula $(A \Rightarrow B) \wedge (B \Rightarrow A)$).

(i) (Strong) Intensional equivalence (or *equality*) "$\backsimeq$" is defined by $< \phi >_{\alpha}^{\beta_1}/g \backsimeq < \psi >_{\alpha}^{\beta_2}/g$ if and only if $\Box(\phi[\beta_1/g(\beta_1)] \equiv \psi[\beta_2/g(\beta_2)])$, with $\mathcal{M} \models_{w,g'} \Box\varphi$ if and only if for all $w' \in \mathcal{W}$, $(w, w') \in \mathcal{R}$ implies $\mathcal{M} \models_{w',g'} \varphi$. From Example 5, we have that $< p_1^1(x) >_x \backsimeq < p_2^1(x) >_x$, that is, "$x$ has been bought" and "x has been sold" are intensionally equivalent, but they have not the same meaning (the concept $I(p_1^1(x)) \in D_1$ is different from $I(p_2^1(x)) \in D_1$).

(ii) Weak intensional equivalence "$\approx$" is defined by $< \phi >_{\alpha}^{\beta_1}/g \approx < \psi >_{\alpha}^{\beta_2}/g$ if and only if $\Diamond\phi[\beta_1/g(\beta_1)] \equiv \Diamond\psi[\beta_2/g(\beta_2)]$. The symbol $\Diamond = \neg\Box\neg$ is the correspondent existential modal operator. This weak equivalence is used for P2P database integration in a number of papers [16, 19–24].

Note that if we want to use the intensional equality in our language, then we need the correspondent operator in intensional algebra $\mathcal{A}_{\text{int}}$ for the "necessity" modal logic operator $\Box$.

This semantics is equivalent to the algebraic semantics for $\mathcal{L}$ in [8] for the case of the conception where intensional entities are considered to be *equal* if and only if they are *necessarily equivalent*. Intensional equality is much stronger that the standard *extensional equality* in the actual world, just because it requires the extensional equality in *all* possible worlds; in fact, if $< \phi >_{\alpha}^{\beta_1}/g \backsimeq < \psi >_{\alpha}^{\beta_1}/g$, then $h(I(< A >_{\alpha}^{\beta_1}/g)) = h(I(< \psi >_{\alpha}^{\beta_2}/g))$ for all extensionalization functions $h \in \mathcal{E}$ (i.e., possible worlds $is^{-1}(h) \in \widetilde{\mathcal{W}}$).

It is easy to verify that the intensional equality means that in every possible world $w \in \widetilde{\mathcal{W}}$ the intensional entities u_1 and u_2 have the same extensions.

Let the logic modal formula $\Box\phi[\beta_1/g(\beta_1)]$, where the assignment g is applied only to free variables in β_1 of a formula ϕ not in the list of variables in $\alpha = (x_1, \ldots, x_n)$, $n \geq 1$, represents an n-ary intensional concept such that $I(\Box\phi[\beta_1/g(\beta_1)]) \in D_n$ and $I(\phi[\beta_1/g(\beta_1)]) = I(< \phi >_{\alpha}^{\beta_1}/g) \in D_n$. Then the extension of this n-ary concept is equal to (here the mapping necess : $D_i \to D_i$ for each $i \geq 0$ is a new operation of the intensional algebra $\mathcal{A}_{\text{int}}$ in Definition 6)

$$\begin{aligned}
&h\left(I\left(\Box\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\\
&= h\left(\text{necess}\left(I\left(\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\right)\\
&= \left\{\left(g'(x_1), \ldots, g'(x_n)\right) \mid \right.\\
&\qquad \left. \mathcal{M} \models_{w,g'} \Box\phi\left[\frac{\beta_1}{g(\beta_1)}\right],\ g' \in \mathcal{D}^{\mathcal{V}}\right\}\\
&= \left\{\left(g'(x_1), \ldots, g'(x_n)\right) \mid g' \in \mathcal{D}^{\mathcal{V}},\ \forall w_1\right.\\
&\qquad \left.\left((w, w_1) \in \mathcal{R} \text{ implies } \mathcal{M} \models_{w_1,g'} \phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right\}\\
&= \bigcap_{h_1 \in \mathcal{E}} h_1\left(I\left(\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right),
\end{aligned} \tag{8}$$

while

$$\begin{aligned}
&h\left(I\left(\Diamond\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\\
&= h\left(I\left(\neg\Box\neg\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\\
&= h\left(\text{neg}\left(\text{necess}\left(I\left(\neg\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\right)\right)\\
&= \mathcal{D}^n \setminus h\left(\text{necess}\left(I\left(\neg\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\right)\\
&= \mathcal{D}^n \setminus \left(\bigcap_{h_1 \in \mathcal{E}} h_1\left(I\left(\neg\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\right)\\
&= \mathcal{D}^n \setminus \left(\bigcap_{h_1 \in \mathcal{E}} h_1\left(neg\left(I\left(\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\right)\right)\\
&= \mathcal{D}^n \setminus \left(\bigcap_{h_1 \in \mathcal{E}} \mathcal{D}^n \setminus h_1\left(I\left(\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right)\right)\\
&= \bigcup_{h_1 \in \mathcal{E}} h_1\left(I\left(\phi\left[\frac{\beta_1}{g(\beta_1)}\right]\right)\right).
\end{aligned} \tag{9}$$

Consequently, the concepts $\Box\phi[\beta_1/g(\beta_1)]$ and $\Diamond\phi[\beta_1/g(\beta_1)]$ are the *built-in* (or rigid) concept as well, whose extensions do not depend on possible worlds.

Thus, two concepts are intensionally *equal*, that is, $\lessdot \phi \gtrdot_{\alpha}^{\beta_1}/g \asymp \lessdot \psi \gtrdot_{\alpha}^{\beta_2}/g$, if and only if $h(I(\phi[\beta_1/g(\beta_1)])) = h(I(\psi[\beta_2/g(\beta_2)]))$ for every h.

Analogously, two concepts are *weakly* equivalent, that is, $\lessdot \phi \gtrdot_{\alpha}^{\beta_1}/g \approx \lessdot \psi \gtrdot_{\alpha}^{\beta_2}/g$ if and only if $h(I(\Diamond\phi[\beta_1/g(\beta_1)])) = h(I(\Diamond\psi[\beta_2/g(\beta_2)]))$.

3. Application to the Intensional FOL without Abstraction Operator

In the case of the intensional FOL defined in Definition 1, without Bealer's intensional abstraction operator $\lessdot\gtrdot$, we obtain the syntax of the standard FOL but with intensional semantics as presented in [15].

Such a FOL has a well-known Tarski's interpretation, defined as follows.

> An interpretation (Tarski) I_T consists in a nonempty domain $\mathscr{D}$ and a mapping that assigns to any predicate letter $p_i^k \in P$ a relation $R = I_T(p_i^k) \subseteq \mathscr{D}^k$, to any functional letter $f_i^k \in F$ a function $I_T(f_i^k) : \mathscr{D}^k \to \mathscr{D}$, or, equivalently, its graph relation $R = I_T(f_i^k) \subseteq \mathscr{D}^{k+1}$ where the $k+1$th column is the resulting function's value, and to each individual constant $c \in F$ one given element $I_T(c) \in \mathscr{D}$.
>
> Consequently, from the intensional point of view, an interpretation of Tarski is a possible world in the Montague's intensional semantics, that is, $w = I_T \in \mathscr{W}$. The correspondent extensionalization function is $h = is(w) = is(I_T)$.
>
> We define the satisfaction of a logic formulae in $\mathscr{L}$ for a given assignment $g : \mathscr{V} \to \mathscr{D}$ inductively, as follows.
>
> If a formula ϕ is an atomic formula $p_i^k(t_1, \ldots, t_k)$, then this assignment g satisfies ϕ if and only if $(g^*(t_1), \ldots, g^*(t_k)) \in I_T(p_i^k)$; g satisfies $\neg\phi$ if and only if it does not satisfy ϕ; g satisfies $\phi \wedge \psi$ iff g satisfies ϕ and g satisfies ψ; g satisfies $(\exists x_i)\phi$ if and only if there exists an assignment $g\prime \in \mathscr{D}^{\mathscr{V}}$ that may differ from g only for the variable $x_i \in \mathscr{V}$, and g' satisfies ϕ.
>
> A formula ϕ is true for a given interpretation I_T if and only if ϕ is satisfied by every assignment $g \in \mathscr{D}^{\mathscr{V}}$. A formula ϕ is valid (i.e., tautology) if and only if ϕ is true for every Tarksi's interpretation $I_T \in \mathfrak{T}_T$. An interpretation I_T is a model of a set of formulae Γ if and only if every formula $\phi \in \Gamma$ is true in this interpretation. We denote by FOL(Γ) the FOL with a set of assumptions Γ, and by $\mathfrak{T}_T(\Gamma)$ the subset of Tarski's interpretations that are models of Γ, with $\mathfrak{T}_T(\emptyset) = \mathfrak{T}_T$. A formula ϕ is said to be a *logical consequence* of Γ, denoted by $\Gamma \Vdash \phi$, if and only if ϕ is true in all interpretations in $\mathfrak{T}_T(\Gamma)$. Thus, $\Vdash \phi$ if and only if ϕ is a tautology.
>
> The basic set of axioms of the FOL are that of the propositional logic with two additional axioms: (A1) $(\forall x)(\phi \Rightarrow \psi) \Rightarrow (\phi \Rightarrow (\forall x)\psi)$ (x does not occur in ϕ and it is not bound in ψ), and (A2) $(\forall x)\phi \Rightarrow \phi[x/t]$, (neither x nor any variable in t occurs bound in ϕ). For the FOL with identity, we need the *proper* axiom (A3) $x_1 \doteq x_2 \Rightarrow (x_1 \doteq x_3 \Rightarrow x_2 \doteq x_3)$.
>
> The inference rules are Modus Ponens and generalization (G) "if ϕ is a theorem and x is not bound in ϕ, then $(\forall x)\phi$ is a theorem."

The standard FOL is considered as an extensional logic because two open sentences with the same tuple of variables $\phi(x_1, \ldots, x_m)$ and $\psi(x_1, \ldots, x_m)$ are equal if and only if they have the *same extension* in a given interpretation I_T, that is, if and only if $I_T^*(\phi(x_1, \ldots, x_m)) = I_T^*(\psi(x_1, \ldots, x_m))$, where I_T^* is the unique extension of I_T to all formulae, as follows.

(1) For a (closed) sentence ϕ/g, we have that $I_T^*(\phi/g) = t$ if and only if g satisfies ϕ, as recursively defined above.

(2) For an open-sentence ϕ with the tuple of free variables $(x_1, \ldots, x_m)$, we have that $I_T^*(\phi(x_1, \ldots, x_m)) =_{\text{def}} \{(g(x_1), \ldots, g(x_m)) \mid g \in \mathscr{D}^{\mathscr{V}} \text{ and } I_T^*(\phi/g) = t\}$.

It is easy to verify that for a formula ϕ with the tuple of free variables $(x_1, \ldots, x_m)$, $I_T^*(\phi(x_1, \ldots, x_m)/g) = t$ if and only if $(g(x_1), \ldots, g(x_m)) \in I_T^*(\phi(x_1, \ldots, x_m))$.

This extensional *equality* of two virtual predicates can be generalized to the extensional *equivalence* when both predicates ϕ, ψ have the same set of free variables but their ordering in the *tuples* of free variables is not identical: such two virtual predicates are equivalent if the extension of the first is equal to the proper permutation of columns of the extension of the second virtual predicate. It is easy to verify that such an extensional equivalence corresponds to the logical equivalence denoted by $\phi \equiv \psi$.

This extensional equivalence between two relations $R_1, R_2 \in \mathfrak{R}$ with the same arity will be denoted by $R_1 \cong R_2$, while the extensional identity will be denoted in the standard way by $R_1 = R_2$.

Let $\mathscr{A}_{\text{FOL}} = (\mathscr{L}, \doteq, \top, \wedge, \neg, \exists)$ be a free syntax algebra for "first-order logic with identity $\doteq$," with the set $\mathscr{L}$ of first-order logic formulae, with $\top$ denoting the tautology formula (the contradiction formula is denoted by $\neg\top$), with the set of variables in $\mathscr{V}$ and the domain of values in $\mathscr{D}$. It is well known that we are able to make the extensional algebraization of the FOL by using the *cylindric* algebras [25] that are the extension of Boolean algebras with a set of binary operators for the FOL identity relations and a set of unary algebraic operators ("projections") for each case of FOL quantification $(\exists x)$. In what follows, we will make an analog extensional algebraization over $\mathfrak{R}$ but by interpretation of the logic conjunction $\wedge$ by a set of *natural join* operators over relations introduced by Codd's relational algebra [9] and [26] as a kind of a predicate calculus whose interpretations are tied to the database.

Corollary 11 (extensional FOL semantics [15]). *Let us define the extensional relational algebra for the FOL by*

$$\mathscr{A}_{\mathfrak{R}} = \left(\mathfrak{R}, R_{=}, \{\langle\,\rangle\}, \{\bowtie_S\}_{S \in \mathscr{P}(\mathbb{N}^2)}, \sim, \{\pi_{-n}\}_{n \in \mathbb{N}}\right), \tag{10}$$

where $\{\langle\,\rangle\} \in \mathfrak{R}$ *is the algebraic value correspondent to the logic truth and* $R_=$ *is the binary relation for extensionally equal elements. We will use "=" for the extensional identity for relations in* $\mathfrak{R}$.

Then, for any Tarski's interpretation I_T *its unique extension to all formulae* $I_T^* : \mathscr{L} \to \mathfrak{R}$ *is also the homomorphism* $I_T^* : \mathscr{A}_{FOL} \to \mathscr{A}_{\mathfrak{R}}$ *from the free syntax FOL algebra into this extensional relational algebra.*

Proof. Directly from definition of the semantics of the operators in $\mathscr{A}_{\mathfrak{R}}$ defined in precedence, let us take the case of conjunction of logic formulae of the definition above where $\varphi(x_i, x_j, x_k, x_l, x_m, y_i, y_j)$ (its tuple of variables is obtained by the method defined in the FOL introduction) is the virtual predicate of the logic formula $\phi(x_i, x_j, x_k, x_l, x_m) \wedge \psi(x_l, y_i, x_j, y_j)$: $I_T^*(\phi \wedge \psi) = I_T^*(\varphi) = \{(g(x_i), g(x_j), g(x_k), g(x_l), g(x_m), g(y_i), g(y_j)) \mid I_T^*(\varphi/g) = t\} = \{(g(x_i), g(x_j), g(x_k), g(x_l), g(x_m), g(y_i), g(y_j)) \mid I_T^*(\phi/g \wedge \psi/g) = t\} = \{(g(x_i), g(x_j), g(x_k), g(x_l), g(x_m), g(y_i), g(y_j)) \mid I_T^*(\phi/g) = t$ and $I_T^*(\phi/g) = t\} = \{(g(x_i), g(x_j), g(x_k), g(x_l), g(x_m), g(y_i), g(y_j)) \mid (g(x_i), g(x_j), g(x_k), g(x_l), g(x_m)) \in I_T^*(\phi)$ and $(g(x_l), g(y_i), g(x_j), g(y_j)) \in I_T^*(\phi)\} = I_T^*(\phi) \bowtie_{\{(4,1),(2,3)\}} I_T^*(\psi)$.

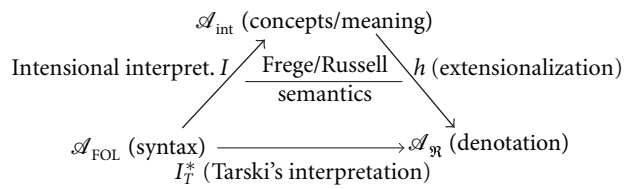
Thus, it is enough to show that $I_T^*(\top) = \{\langle\,\rangle\}$ is also valid, and $I_T^*(\neg\top) = \emptyset$. The first property comes from the fact that $\top$ is a tautology, thus satisfied by every assignment g, that is, it is true, that is, $I_T^*(\top) = t$ (and t is equal to the empty tuple $\{\langle\,\rangle\}$). The second property comes from the fact that $I_T^*(\neg\top) = \sim (I_T^*(\top)) = \sim (\{\langle\,\rangle\}) = \mathscr{D}^0 \setminus \{\langle\,\rangle\} = \{\langle\,\rangle\} \setminus \{\langle\,\rangle\} = \emptyset$. That is, the tautology and the contradiction have the true and false logic value, respectively, in $\mathfrak{R}$.

We have also that $I_T^*(\doteq (x, y)) = I_T(\doteq) = R_=$ for every interpretation I_T because $\doteq$ is the built-in binary predicate, that is, with the same extension in every Tarski's interpretation.

Consequently, the mapping $I_T^* : (\mathscr{L}, \doteq, \top, \wedge, \neg, \exists) \to \mathscr{A}_{\mathfrak{R}}$ is a homomorphism that represents the extensional Tarskian semantics of the FOL. □

Consequently, we obtain the following Intensional/extensional FOL semantics [15].

For any Tarski's interpretation I_T of the FOL, the following diagram of homomorphisms commutes.



(11)

where $h = is(w)$ and $w = I_T \in \mathscr{W}$ is the explicit possible world (extensional Tarski's interpretation).

This homomorphic diagram formally expresses the fusion of Frege's and Russell's semantics [27–29] of meaning and denotation of the FOL language and renders mathematically correct the definition of what we call an "intuitive notion of intensionality," in terms of which a language is intensional if denotation is distinguished from sense: that is, if both denotation and sense are ascribed to its expressions. This notion is simply adopted from Frege's contribution (without its infinite sense-hierarchy, avoided by Russell's approach where there is only one meaning relation, one fundamental relation between words and things, here represented by one fixed intensional interpretation I), where the sense contains mode of presentation (here described algebraically as an algebra of concepts (intensions) $\mathscr{A}_{\text{int}}$), and where sense determines denotation for any given extensionalization function h (correspondent to a given Traski's interpretation I_T). More about the relationships between Frege's and Russell's theories of meaning may be found in the Chapter 7, "Extensionality and Meaning", in [18].

As noted by Gottlob Frege and Rudolf Carnap (he uses terms Intension/extension in the place of Frege's terms sense/denotation [30]), the two logic formulae with the same denotation (i.e., the same extension for a given Tarski's interpretation I_T) need not have the same sense (intension), thus such codenotational expressions are not *substitutable* in general.

In fact there is exactly *one* sense (meaning) of a given logic formula in $\mathscr{L}$, defined by the uniquely fixed intensional interpretation I, and *a set* of possible denotations (extensions) each determined by a given Tarski's interpretation of the FOL as follows from Definition 4:

$$\mathscr{L} \xrightarrow{I} \mathscr{D} \Longrightarrow_{h=is(I_T)\&I_T \in \mathscr{W} = \Im_T(\Gamma)} \mathfrak{R}. \tag{12}$$

Often "intension" has been used exclusively in connection with possible worlds semantics; however, here we use (as many others; as Bealer for example) "intension" in a more wide sense, that is, as an *algebraic expression* in the intensional algebra of meanings (concepts) $\mathscr{A}_{\text{int}}$, which represents the structural composition of more complex concepts (meanings) from the given set of atomic meanings. Consequently, not only the denotation (extension) is compositional, but also the meaning (intension) is compositional.

4. Conclusion

Semantics is a theory concerning the fundamental relations between words and things. In Tarskian semantics of the FOL, one defines what it takes for a sentence in a language to be truely relative to a model. This puts one in a position to define what it takes for a sentence in a language to be valid. Tarskian semantics often proves quite useful in logic. Despite this, Tarskian semantics neglects meaning, as if truth in language were autonomous. Because of that the Tarskian theory of truth becomes inessential to the semantics for more expressive logics, or more "natural" languages.

Both Montague's and Bealer's approaches were useful for this investigation of the intensional FOL with intensional abstraction operator, but the first is not adequate and explains why we adopted two-step intensional semantics (intensional interpretation with the set of extensionalization functions).

At the end of this work, we defined an extensional algebra for the FOL (different from standard cylindric algebras) and

the commutative homomorphic diagram that expresses the generalization of the Tarskian theory of truth for the FOL into the Frege/Russell's theory of meaning.

References

[1] J. Pustejovsky and B. Boguraev, "Lexical knowledge representation and natural language processing," *Artificial Intelligence*, vol. 63, no. 1-2, pp. 193–223, 1993.

[2] M. Fitting, "First-order intensional logic," *Annals of Pure and Applied Logic*, vol. 127, no. 1–3, pp. 171–193, 2004.

[3] T. Braüner, "Adding intensional machinery to hybrid logic," *Journal of Logic and Computation*, vol. 18, no. 4, pp. 631–648, 2008.

[4] T. Braüner and S. Ghilardi, "First-order modal logic," in *Handbook of Modal Logic*, pp. 549–620, Elsevier, 2007.

[5] S. Bond and M. Denecker, "I-logic: an intensional logic of informations," in *Proceedings of the 19th Belgian-Dutch Conference on Artificial Intelligence*, pp. 49–56, Utrecht, The Netherlands, November 2007.

[6] M. A. Orgun and W. W. Wadge, "Towards a unified theory of intensional logic programming," *The Journal of Logic Programming*, vol. 13, no. 4, pp. 413–440, 1992.

[7] G. Bealer, "Universals," *The Journal of Philosophy*, vol. 90, no. 1, pp. 5–32, 1993.

[8] G. Bealer, "Theories of properties, relations, and propositions," *The Journal of Philosophy*, vol. 76, no. 11, pp. 634–648, 1979.

[9] E. F. Codd, "Relational completeness of data base sublanguages," in *Data Base Systems: Courant Computer Science Symposia Series 6*, Prentice Hall, Englewood Cliffs, NJ, USA, 1972.

[10] D. K. Lewis, *On the Plurality of Worlds*, Blackwell, Oxford, UK, 1986.

[11] R. Stalnaker, *Inquiry*, The MIT Press, Cambridge, Mass, USA, 1984.

[12] R. Montague, "Universal grammar," *Theoria*, vol. 36, no. 3, pp. 373–398, 1970.

[13] R. Montague, "The proper treatment of quantification in ordinary English," in *Approaches to Natural Language*, J. Hintikka, P. Suppes, J. M. E. Moravcsik et al., Eds., pp. 221–242, Reidel, Dordrecht, The Netherlands, 1973.

[14] R. Montague, *Formal Philosophy. Selected Papers of Richard Montague*, Yale University Press, London, UK, 1974.

[15] Z. Majkić, "First-order logic: modality andintensionality," http://arxiv.org/abs/1103.0680v1.

[16] Z. Majkić, "Intensional first-order logic for P2P database systems," in *Journal on Data Semantics 12*, vol. 5480 of *Lecture Notes in Computer Science*, pp. 131–152, Springer, Berlin, Germany, 2009.

[17] Z. Majkić, "Intensional semantics for RDF data structures," in *Proceedings of the 12th International Symposium on Database Engineering & Applications Systems (IDEAS '08)*, pp. 69–77, Coimbra, Portugal, September, 2008.

[18] G. Bealer, *Quality and Concept*, Oxford University Press, New York, NY, USA, 1982.

[19] Z. Majkić, "Weakly-coupled ontology integration of P2P database systems," in *Proceedings of the 1st International Workshop on Peer-to-Peer Knowledge Management (P2PKM '04)*, Boston, Mass, USA, August 2004.

[20] Z. Majkić, "Intensional P2P mapping between RDF ontologies," in *Proceedings of the 6th International Conference on Web Information Systems (WISE '05)*, M. Kitsuregawa, Ed., vol. 3806 of *Lecture Notes in Computer Science*, pp. 592–594, New York, NY, USA, November 2005.

[21] Z. Majkić, "Intensional semantics for P2P data integration," in *Journal on Data Semantics 6*, vol. 4090 of *Lecture Notes in Computer Science*, Special Issue on 'Emergent Semantics', pp. 47–66, 2006.

[22] Z. Majkić, "Non omniscient intensional contextual reasoning for query-agents in P2P systems," in *Proceedings of the 3rd Indian International Conference on Artificial Intelligence (IICAI '07)*, Pune, India, December 2007.

[23] Z. Majkić, "Coalgebraic specification of query computation in intensional P2P database systems," in *Proceedings of the International Conference on Theoretical and Mathematical Foundations of Computer Science (TMFCS '08)*, pp. 14–23, Orlando, Fla, USA, July 2008.

[24] Z. Majkić, "RDF view-based interoperability in intensional FOL for Peer-to-Peer database systems," in *Proceedings of the International Conference on Enterprise Information Systems and Web Technologies (EISWT '08)*, pp. 88–96, Orlando, Fla, USA, July 2008.

[25] L. Henkin, J. D. Monk, and A. Tarski, *Cylindic Algebras I*, North-Holland, Amsterdam, The Netherlands, 1971.

[26] A. Pirotte, "A precise definition of basic relational notions and of the relational algebra," *ACM SIGMOD Record*, vol. 13, no. 1, pp. 30–45, 1982.

[27] G. Frege, "Über Sinn und Bedeutung," *Zeitschrift für Philosophie und philosophische Kritik*, vol. 100, pp. 22–50, 1892.

[28] B. Russell, "On Denoting," in *Logic and Knowledge*, vol. 14 of *Mind*, Reprinted in Russell, pp. 479–493, 1905.

[29] A. N. Whitehead and B. Russell, *Principia Mathematica*, vol. 1, Cambridge, Mass, USA, 1910.

[30] R. Carnap, *Meaning and Necessity*, Chicago, Ill, USA, 1947.

Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2012, Article ID 674832, 19 pages
doi:10.1155/2012/674832

Research Article

Basin Hopping as a General and Versatile Optimization Framework for the Characterization of Biological Macromolecules

Brian Olson,[1] Irina Hashmi,[1] Kevin Molloy,[1] and Amarda Shehu[1,2]

[1] *Department of Computer Science, George Mason University, Fairfax, VA 22030, USA*
[2] *Department of Bioengineering, George Mason University, Fairfax, VA 22030, USA*

Correspondence should be addressed to Amarda Shehu, amarda@gmu.edu

Received 29 June 2012; Revised 23 September 2012; Accepted 19 October 2012

Academic Editor: Zhiyuan Luo

Since its introduction, the basin hopping (BH) framework has proven useful for hard nonlinear optimization problems with multiple variables and modalities. Applications span a wide range, from packing problems in geometry to characterization of molecular states in statistical physics. BH is seeing a reemergence in computational structural biology due to its ability to obtain a coarse-grained representation of the protein energy surface in terms of local minima. In this paper, we show that the BH framework is general and versatile, allowing to address problems related to the characterization of protein structure, assembly, and motion due to its fundamental ability to sample minima in a high-dimensional variable space. We show how specific implementations of the main components in BH yield algorithmic realizations that attain state-of-the-art results in the context of ab initio protein structure prediction and rigid protein-protein docking. We also show that BH can map intermediate minima related with motions connecting diverse stable functionally relevant states in a protein molecule, thus serving as a first step towards the characterization of transition trajectories connecting these states.

1. Introduction

Global optimization is an objective of many disciplines, both in academic and industrial settings [1, 2]. Characterization of complex systems often poses very hard global optimization problems with many variables [3, 4]. Algorithms that target such problems largely build on or combine four main approaches: deterministic, stochastic, heuristic, and smoothing [3, 5–7]. All these algorithms are challenged by systems where the variable space contains multiple distinct minima. While most algorithms can efficiently find a minimum, not all can feasibly locate the global minimum.

Some of the most successful applications of global optimization algorithms on characterizing physical and biological systems build on the stochastic Monte Carlo (MC) procedure and its Metropolis variant [8]. For instance, simulated annealing is one of the most widely used algorithms for finding the global minimum of a multivariable function for different complex systems [4, 9, 10]. Adaptations that build on deterministic and stochastic numerical procedures, such as molecular dynamics (MD) and MC, are abundant in computational biology for the structural characterization of biological macromolecules (cf. [11, 12]).

Basin hopping (BH) is a global optimization framework that is particularly suited for multivariable multimodal optimization problems [13], and it is our thesis in this paper that BH is an effective framework for the characterization of biological macromolecules. The basic BH framework is well studied and understood, but modifications to its core components are necessary for application to complex biological systems. In what follows, we first summarize the basic BH framework and some of its salient properties before proceeding to identify modifications necessary for application to biological macromolecules.

BH combines heuristic procedures with local searches to enhance its exploration of the given variable space, conducted as a series of perturbations followed by local optimization. As shown in pseudocode in Algorithm 1,

```
(1) i ← 0
(2) X_i ← random initial point in variable space
(3) Y_i ← LOCALSEARCH (X_i)
(4) while STOP not satisfied do
(5)     X_{i+1} ← PERTURB (Y_i)
(6)     Y_{i+1} ← LOCALSEARCH (X_{i+1})
(7)     if f(Y_{i+1}) < f(Y_i) then
(8)         i ← i + 1
```

Algorithm 1: Basic BH framework in pseudocode.

the framework can be described in terms of a local search procedure LOCALSEARCH that maps a point X_i in variable space to its nearest minimum Y_i, a perturbation move PERTURB that modifies a current minimum Y_i to obtain a new point X_{i+1} in variable space, and a stopping criterion STOP that terminates these repeated applications of a structural perturbation followed by a local optimization. The repeated applications result in a trajectory of local minima Y_i. As shown in Algorithm 1, only the lowest minimum needs to be retained in memory when seeking the global minimum of some function f. It is important to note that Algorithm 1 shows a specific realization of the BH framework, known as monotonic BH (MBH), where the current minimum is not accepted if it does not lower the lowest value obtained for the function f so far. In this case, another perturbation is attempted in order to obtain a new starting point for the local optimization that follows.

While this basic framework is easy to describe and employ for global optimization, effective implementations exploit specific domain expertise about the system at hand [14–19]. Heuristics are designed based on specific system knowledge to implement an effective perturbation component. Domain-specific expertise is also employed for an effective implementation of the local search component. The stopping criterion is often implemented in terms of a maximum number of function evaluations or in terms of no improvements over a window of the last sampled minima. It is important to note that the stochasticity in BH is mainly due to the implementation of the perturbation component, which seeks to take the exploration out of the current local minimum. The local optimization component, on the other hand, can employ deterministic numerical techniques to locate the local minimum with arbitrary accuracy [20].

The core advantage of the BH framework over a multistart method that essentially samples local minima at random is that BH moves between adjacent local minima in the variable space. This strategy is more effective when exploring high-dimensional variable spaces associated with complex physical systems, where the addition of new dimensions can result in an exponential increase in the number of minima in the space [21]. The adjacency is a result of a deep connection between the perturbation and local optimization. Despite the application setting, a good general rule is for the perturbation to preserve some structural characteristics of the local minimum Y_i it is disrupting to obtain a new starting point X_{i+1} for the next application of the local optimization.

If the magnitude of the perturbation jump in the variable space, measured through some distance function $d(Y_i, X_{i+1})$, is small, then X_{i+1} may remain in the basin of attraction of Y_i, and the local optimization will bring X_{i+1} back to Y_i. On the other end of the spectrum, the perturbation can completely disrupt Y_i and obtain an X_{i+1} that could have essentially been obtained at random. While the local optimization will yield a new local minimum $Y_{i+1} \neq Y_i$, the BH will degenerate to a multistart method in this case. Different studies have shown that the perturbation needs to preserve some of the structure of the current minimum for BH to be more effective than the multistart method [20, 22]. It is the careful implementation of the perturbation component that allows BH to organize the local minima it samples according to an adjacency relationship [21].

The BH framework is sometimes referred to as a funnel-descent method, because its core behavior of iterating over adjacent local minima has turned out to be an effective optimization strategy for functions with a funnel landscape [21]. The generality of the framework and its ease of adaption for different systems has resulted in diverse applications, which span from geometry problems, such as packing circles in circular containers [20], to statistical physics problems of characterizing low-energy states of small atomic clusters [3].

The BH framework originated in the computational biology community dating back to the pioneering work of Wales [3], where the objective was to characterize the minima of the Lennard-Jones energy function in small atomic clusters. The term basin hopping was coined in this work, though to an extent, the stated motivation for the BH framework was from related optimization algorithms in the evolutionary search community. In fact, BH can be viewed as a special case of Iterated Local Search, which is popular for solving discrete combinatorial optimization problems [23]. An algorithmic realization of the BH framework was available prior to the work of Wales, most notably in Scheraga's MC with minimization algorithm [9, 24].

The BH framework is particularly suited to deal with molecular spaces, where the function sought for optimization is a complex nonconvex potential energy function summing over the interactions among atoms in a 3-dimensional molecular structure. The global minimum of the function corresponds to the structural state of the molecule that is most stable under equilibrium conditions and so relevant for biological activity. Structural characterization of the biologically active (native) state of biological macromolecules is

an important problem in computational structural biology. A grand-standing challenge nowadays is to characterize such states for protein molecules, which are central in many chemical pathways in the cell and are the focus of this paper.

Proteins are complex systems with hundreds to thousands of atoms. These atoms are organized in amino-acid building blocks which connect serially to form a polypeptide chain (the N-terminus of one amino acid connects to the C-terminus of the other to form a peptide). Figure 1(a) shows a short polypeptide chain. Depending on the representation employed, a spatial arrangement of the atoms that constitute a polypeptide chain, also referred to as a conformation, may require the specification of a prohibitive number of variables. A popular representation in computational structural biology employs only the angles shown in Figure 1(a). These angles can be used to define the variable space, as their modification gives rise to different conformations.

The variable, or conformational, space of a polypeptide chain is associated with a funnel-like energy surface [25, 26]. The size and ruggedness of this surface, illustrated in Figure 1(b), are the primary reasons why obtaining structural information on native state of a protein polypeptide chain based on the chain's amino-acid sequence alone is an outstanding challenge in computational structural biology [27]. Meeting this challenge, often known as ab initio protein structure prediction, is needed, however, to close the gap between the wealth of protein sequence data and the scarce information on their native structures. Obtaining structural information ab initio promises to elucidate the structure-function relationship and advance structure-driven studies and applications on protein molecules [28–30].

The funnel-like but rugged energy surface of protein molecules seems suitable for the BH framework. In general, it is challenging to locate the global minimum in this surface and so elucidate the native structure of a protein. One of the main reasons relates to imperfect modeling. The energy functions currently available to probe the protein energy surface are semiempirical and contain inherent errors [31]. Due to the specific process undertaken in computational chemistry to design such functions, the actual global minimum of a designed protein energy function may deviate significantly from the true global minimum (the native structure obtained by experiment in the wet laboratory). Studies report deviations in the 2–4 Å range [32]. Due to these deviations, computational approaches that aim to obtain a broad view of the energy surface are more appropriate, particularly if they are to be followed by detailed heavy-duty optimization techniques on select conformations.

A common strategy among protocols for ab-initio protein structure prediction is the sampling of a large number of low-energy conformations. These are end points of many independent MD or MC trajectories optimizing some chosen energy function [28, 33–39]. Alternatively, the trajectories can be integrated in a tree to better control the exploration and use online analysis to bias the tree away from high-energy oversampled regions [40, 41]. The conformations are then grouped by structural similarity to reveal local minima from which it is worth continuing the exploration at higher representational detail. The goal then becomes obtaining convergence to a region of the space that can be predicted to represent the native state.

In the context of ab-initio protein structure prediction, BH can be employed to explicitly sample local minima in the protein energy surface. At a superficial level, this would require the retainment of an ensemble of local minima and not just the current one. In addition, while the pseudocode in Algorithm 1 shows a simple realization of the BH framework, MBH, applications of BH on molecular spaces often make probabilistic decisions on whether to accept a current minimum. Procedurally, the framework still consists of repeated applications of a structural perturbation followed by an energy minimization. However, a Metropolis criterion [42] biases the sampling of local minima towards lower energy ones over time. Essentially, the decision to accept Y_{i+1} is made with probability $\exp(-[E(Y_{i+1})-E(Y_i)]/[K_BT])$, where E refers to the energy function, K_B is the Boltzmann constant, and T is temperature. Temperature does not need a physical meaning, as its main role is to scale the height of an energy barrier.

The appeal of the BH framework is that it transforms the protein energy surface into a collection of interpenetrating staircases, as illustrated in Figure 1(c). A succinct discrete representation is obtained for this surface in terms of local minima. It is important to note that BH does not modify the energy surface in any way. Instead, it projects each point (conformation) to its closest local minimum to effectively reveal a map of the energy surface in terms of local minima. The details of the energy surface between local minima are lost, but this degree of resolution is still very useful for a structural characterization of protein molecules.

Given its ease of implementation, BH is starting to gain popularity as an optimization framework for biological systems. Current applications of BH for structural characterization of biological molecules essentially differ in the specific implementations for the perturbation and local optimization components. Local optimization, for instance, is implemented as gradient descent or Metropolis MC at low temperature, whereas the perturbation component, on the other hand, directly modifies atomic coordinates in existing work. These implementation choices have allowed BH algorithms to capture local minima of small atomic clusters and even map energy surfaces of polyalanines and other small proteins [14, 32, 43, 44]. However, applications to structure prediction [45] have been limited to small proteins, mainly because representation of conformations through atomic coordinates results in a prohibitive variable space. In particular, the BH algorithm in [45] succeeds in locating conformations closer to the experimentally determined native structure than MD with simulated annealing, but its efficiency drops on sequences longer than 75 amino acids.

In this paper, we show that BH is a useful framework for structural characterization beyond structure prediction. We recognize that BH is general and can be employed to map the equilibrium conformational space of a biological system. For instance, we show that with suitable modifications to the perturbation and local optimization components, BH can be applied to protein-protein docking to reveal native

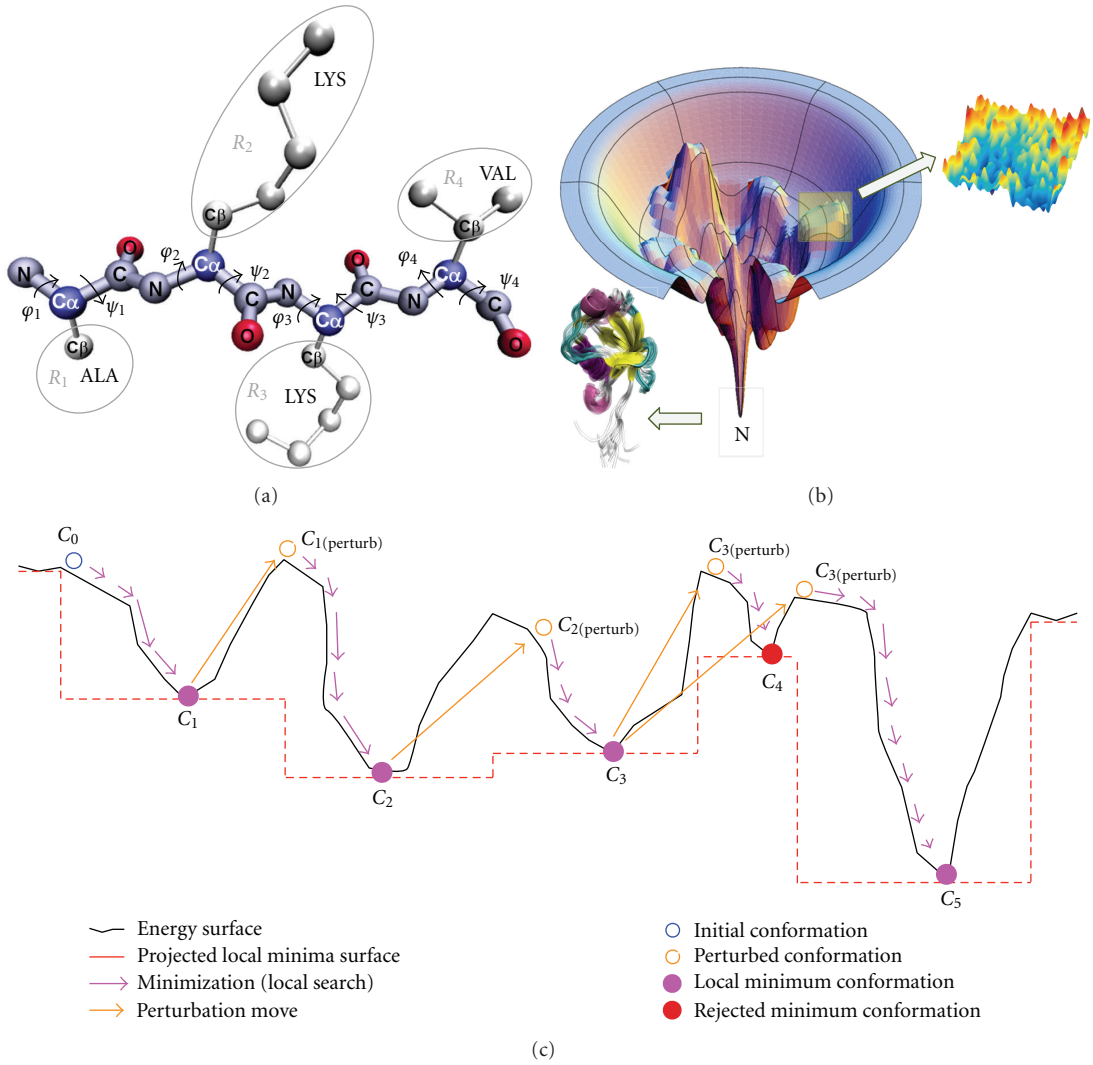


FIGURE 1: (a) A short polypeptide chain of 4 amino acids, alanine, lysine, lysine, and valine, is shown. The backbone atoms shared by all amino acids are N, C_α, C, and O. Side-chain atoms unique to 20 types of amino acids are in gray. The backbone (ϕ, ψ) dihedral angles are annotated over the chain. (b) A model energy surface is illustrated, adapted from [25]. The surface is funnel-like but rugged. The native state at the bottom is denoted by N. Conformations associated with it (obtained from experiment) are illustrated for a particular protein molecule. (c) The BH framework essentially converts the function into a stepwise one. The perturbation and local optimization components are illustrated here with differently colored arrows. A minimum is shown here which fails the Metropolis criterion and is thus not accepted, prompting a new perturbation move.

lowest-energy configurations of protein molecules resulting from the assembly of various polypeptide chains. Specifically, in the context of ab-initio protein structure prediction, we show that implementations of the main components in BH that employ domain-specific knowledge result in increased efficiency and allow application to longer protein chains.

We also show that the ability of BH to provide a map of the energy surface in terms of minima is useful not only when the goal is to locate the global minimum (whether that minimum corresponds to the native structure of one protein polypeptide chain or of a complex resulting from assembly of multiple chains), but also when the objective is to characterize proteins with more than one functionally relevant state. Such proteins are abundant in biology as effective biological machines that can tune their biological function through molecular motions [46–49]. A map of the minima surrounding the functionally relevant states is useful for understanding how the protein hops between minima in transition trajectories connecting these states [33, 49, 50].

The presentation in this paper of BH as a general, versatile framework builds over our recent work on ab-initio structure prediction and rigid protein-protein docking [22, 51]. In particular, in the context of structure prediction, we show that employment of the molecular fragment replacement technique allows BH to efficiently capture the native structure. In the context of protein-protein docking, we incorporate geometric hashing to efficiently obtain structural perturbations of a dimeric configuration.

Additional information from evolutionary sequence analysis allows restricting the variable space. Finally, we provide here a proof-of-concept demonstration that BH can be applied to understand the connectivity between functionally relevant states in a protein in terms of the minima surrounding these states. Obtaining a view of minima in the equilibrium conformational space of a protein molecule is the first step into elucidating motions and transition trajectories that take a protein between the states it uses for biological function.

2. Methods

The basic BH framework was showcased in Algorithm 1 in Section 1. Our algorithmic treatment in this section focuses on modifications to the basic components of BH which allow its application to the three different problems on which we focus in this paper. As described in Section 1, two modifications that allow application to these problems concern accepting a newly obtained local minimum according to the Metropolis criterion (unlike the basic MBH algorithm) and adding that minimum to a growing ensemble of BH-obtained local minima (unlike recording only the last one as in the basic MBH framework). The description of BH below is organized according to the three different applications showcased in this paper in Sections 2.1, 2.2, and 2.3, respectively. The treatment of BH in each application is limited to description of four main components: (1) representation of the system being modeled, which allows defining the variable space; (2) description of the energy function being optimized by BH; (3) implementation of the structural perturbation move; and (4) implementation of the local search procedure for the local optimization.

2.1. BH for Sampling Decoy Conformations for Ab Initio Protein. As described above, the BH framework can be employed to obtain a broad view of the energy surface in terms of low-energy local minima. This can be done efficiently at a coarse-grained level of detail, employing an energy function that sacrifices detail and some accuracy to save computational time. The sampled conformations corresponding to the local minima are low-energy decoy conformations, which can then be fed to any structure prediction protocol for further analysis and refinement of select conformations with dedicated computational resources. The refinement will allow adding further detail, discriminating between decoy conformations, and making a prediction on which refined conformation can be considered to represent the native structure.

2.1.1. Employed Representation. As illustrated in Figure 1(a), a polypeptide chain of n amino acids contains $2n$ backbone (ϕ, ψ) dihedral angles. Our representation of a protein conformation employs only these angles, which constitute the variable space. Side chains are sacrificed, as any structure prediction protocol can pack them as part of the ensuing refinement of decoy conformations [52]. The representation here is essentially the idealized geometry model, which fixes bond lengths and angles to idealized (native) values. Forward kinematics allows computing Cartesian coordinates of the backbone atoms (on which the energy function described below operates) from the ϕ, ψ angles in the representation [53].

2.1.2. Energy Function. The energy function is a modification of the associative memory Hamiltonian with water (AMW) [54]. This function has been used previously by us and others in the context of ab-initio structure prediction [40, 41, 55–57]. AMW sums nonlocal terms (local interactions are kept at ideal values in the idealized geometry model): $E_{\text{AMW}} = E_{\text{Lennard-Jones}} + E_{\text{H-Bond}} + E_{\text{contact}} + E_{\text{burial}} + E_{\text{water}} + E_{\text{Rg}}$. The $E_{\text{Lennard-Jones}}$ term is implemented after the 12–6 Lennard-Jones potential in AMBER9 [58] but allows a soft penetration of van der Waals spheres. The $E_{\text{H-Bond}}$ term allows modeling hydrogen bonds and is implemented as in [59]. The other terms, E_{contact}, E_{burial}, and E_{water}, allow formation of nonlocal contacts, a hydrophobic core, and water-mediated interactions, and are implemented as in [39].

The listed energy terms of E_{AMW} sum over pairwise interactions. For instance, the 12–6 functional form of the Lennard-Jones term is $-4\epsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}]$, where ϵ_{ij} is a constant characteristic of the types of atoms at positions i and j, σ_{ij} is the average diameter of the atoms, and r_{ij} is their distance. This functional form illustrates the quadratic running time of a typical energy function modeling pairwise interactions. More importantly, the terms summed together in an energy function are competing; minima of one term are obtained by suboptima of the other. This competition is known as frustration and refers to the fact that slight changes in atomic positions may lower the value of one term but increase that of another term. The result of summing competing terms in an energy function is a complex multimodal function, whose optimization is nontrivial. More details on the functional form of the other terms of the AMW energy function can be found in [33, 54].

2.1.3. Implementation of Structural Perturbation. The realization of the BH framework we describe here hops between two conformations representing two consecutive minima C_i and C_{i+1} through an intermediate $C_{\text{perturb},i}$ conformation. The perturbation modifies C_i to obtain a higher-energy conformation $C_{\text{perturb},i}$ to escape the current minimum. Essentially, 6 backbone dihedral angles of a fragment of the polypeptide chain associated with three consecutive amino acids in the current conformation C_i are modified simultaneously. This process is referred to as the molecular fragment replacement technique, because it allows replacing the current configuration (in terms of angles) of a selected fragment with another fragment configuration [60].

Molecular Fragment Replacement. The fragment replacement technique has allowed ab-initio structure prediction methods to make great advancements [28, 34–37]. Its key advantage is that it allows obtaining physically realistic modifications if the fragment configurations are sampled from a library of actual native structures obtained in the wet laboratory. The basic idea is that a subset of nonredundant

protein structures are obtained from the Protein Data Bank [61], and configurations of all fragments that can be defined for k consecutive amino acids are excised from these structures and stored in a library. We direct the reader to [28, 41] for a detailed description of how the library is constructed. In this work, we employ a fragment of length 3 rather than a longer fragment, so that the magnitude of the jump resulting from the fragment replacement in variable space is limited.

The perturbation component is implemented as follows. Given a conformation C_i, a fragment of length 3 is selected at random over the polypeptide chain ($n - 2$ fragments can be defined with overlap over a chain of n amino acids). Once the fragment is selected, a configuration for that fragment is then sampled at random over those available for the fragment from the fragment configuration library. The replacement of the angles of the fragment in C_i with those of the configuration obtained from library results in $C_{\text{perturb},i}$.

Since low-energy conformations tend to be compact and leave little room for movement without raising energy (a concept known as frustration in protein biophysics), this implementation of the perturbation component is sufficient to obtain a high-energy conformation through which to escape the current local minimum. Additionally, $C_{\text{perturb},i}$ will share nearly all of its local structural features with C_i, but the new conformation will have a higher energy and a different overall global structure. We note that the first conformation to initiate BH is obtained after $n - 2$ fragment configuration replacements over an extended conformation.

2.1.4. Implementation of Local Optimization. Our implementation of the local optimization conducts a series of modifications starting from $C_{\text{perturb},i}$ to reach a new minimum C_{i+1}. While numerical techniques can be used here, they tend to be inefficient [45]. We employ instead a greedy search, which essentially attempts a maximum of m consecutive fragment replacements (as described above for the perturbation component) until k consecutive attempts fail to lower energy. The resulting C_{i+1} conformation is added to the trajectory according to the Metropolis criterion based on the energetic difference with C_i.

Our implementation of the local optimization is probabilistic due to the fragment replacement technique. Moreover, the true bottom of a current basin may not be found. A working definition of a local minimum is employed instead in terms of the parameter k. Finding true local minima in the energy surface can be computationally intensive while unnecessary. For instance, analysis of the AMW surface in related work in [22] shows that the native structure is near but not at a minimum. In addition, the results in Section 3 make the case that a working definition of a local minimum is sufficient to discover near-native conformations.

2.2. BH for Sampling Decoy Configurations for Rigid Protein-Protein Docking. In this application, the native structures of two protein polypeptide chains (referred to as monomers) are known atomic coordinates obtained for each of the chains from experiment or structure prediction protocols. The objective is to find the native quaternary structure that brings the two monomers together. The assumption here is that the monomers do not change structure upon docking but bind rigidly with each other. Under this assumption, the objective is to find the spatial arrangement that brings one monomeric structure over the other and results in a dimeric configuration of lowest energy.

2.2.1. Employed Representation. In rigid docking, the only variables of interest are those that allow representing a spatial arrangement of one monomeric structure over another. A natural way to do so is through rigid-body transformations, which can be represented as vectors of 6 variables (3 for translation and 3 for rotation in 3-dimensional space). Hence, the variable space here is the 6-dimensional SE(3) space consisting of rigid-body motions or transformations.

The variable space we consider here is not the entire SE(3) but is constrained to rigid-body motions that align geometrically complementary and evolutionary conserved regions of the molecular spaces associated with each of the monomers. This builds upon earlier work by us on rigid docking which makes use of geometric hashing [51, 62]. Geometric hashing is a popular technique that essentially discretizes the space of rigid-body transformations by defining these transformations as alignments of geometrically complementary regions on monomeric molecular surfaces [63–66]. In recent work [51, 62] we show that the number of regions relevant for alignment can be further reduced by focusing on regions with high evolutionary conservation. Such regions are often found to be on contact interfaces [67].

While details of the process through which rigid-body transformations are defined are available in previous work [51, 62], we provide here a brief summary. The Connolly representation is first obtained for each monomeric surface [68]. The representation stores geometrical information for points on the surface, including whether the point represents a convex, saddle, or concave region. The representation is made less dense by only storing key locations on the molecular surface, known as critical points [69]. Triangles can be defined over these points. Associating evolutionary information with a critical point (through an analysis of related biological sequences [67]) allows focusing on triangles with high sequence conservation. We refer to these as active triangles. Once two geometrically complementary (e.g., concave with convex) active triangles T_A and T_B are obtained (from the molecular surfaces of monomers A and B, resp.), a rigid body transformation is easily defined as the one that aligns the local coordinate frame associated with T_B over that associated with T_A.

2.2.2. Energy Function. Each rigid-body motion can be represented as a transformation, a vector of 3 translation and 3 rotation components (details below) that when applied to the moving monomer (one monomer is designated as moving and the other as reference or base) move that monomer in space and bring it over the reference monomer. Atomic coordinates are then obtained for the resulting dimeric configuration, which can now be evaluated in terms

of the interaction energy. The energy function we employ combines three nonlocal terms useful for contact interfaces: $E = E_{\text{VdW}} + E_{\text{electrostatic}} + E_{\text{hydrogen-bonding}}$. The first term implements the standard 12–6 Lennard-Jones potential as in the CHARMM force field [70]. The electrostatic term implements Coulomb's law, also as in the CHARMM force field [70]. The hydrogen-bonding term is calculated as in [71] through the 12–10 hydrogen potential: $E_{\text{hydrogen-bonding}} = 5 \times [(r_0/d_{ij})^{12} - 6 \times (r_0/d_{ij})^{10}]$, where d_{ij} is the distance between acceptor and donor atoms i and j, and $r_0 = 2.9$ Å is the optimal distance for hydrogen bonding. Energy is computed only for the contact interface, which is defined over pairs of atoms in one monomer in contact with the atoms in the other monomer. Two atoms are in contact if their Euclidean distance is not higher than 4.0 Å.

2.2.3. Implementation of Structural Perturbation. The exposition above describes that a rigid-body motion is obtained by aligning an active triangle T_B on the surface of monomer B with a geometrically complementary active triangle T_A on the surface of the base monomer A. Let the current minimum C_i be the configuration corresponding to the transformation aligning T_B with T_A. In other words, the contact interface in C_i is that obtained by aligning T_B with T_A. As described in Section 1, an effective structural perturbation needs to preserve the adjacency relationship. For this reason, an effective perturbation in this context needs to modify the contact interface in C_i but limit the magnitude of the perturbation. The implementation we pursue here seeks a new pair of triangles, T'_A and T'_B, to perturb C_i and obtain $C_{\text{perturb},i}$. In order to limit the magnitude of the perturbation and preserve some of the contact interface of C_i in $C_{\text{perturb},i}$, T'_A needs to be close to T_A, and T'_B needs to be close to T_B.

This is implemented as follows. The molecular surface of each monomer is precomputed and represented in terms of a finite list of active triangles. The center of mass of each triangle is computed, and reverse indexing is used in order to sample a triangle T'_A and a triangle T'_B whose center of mass is within d Å of the center of mass of triangles T_A and T_B, respectively. The process repeats until a pair T'_A and T'_B are found which are geometrically complementary. A new rigid-body transformation aligning T'_B with T'_A is then defined, resulting in the perturbed configuration $C_{\text{perturb},i}$. Sampling in a d-radius neighborhood allows controlling and limiting the extent to which $C_{\text{perturb},i}$ perturbs the structural features of C_i (in this context, the contact interface).

2.2.4. Implementation of Local Optimization. As in the realization of the BH framework for protein structure prediction, the local optimization here also attempts at most m structural modifications starting with $C_{\text{perturb},i}$. The optimization terminates early if k consecutive modifications fail to lower energy. A naive implementation of the local optimization could employ the same structural modifications as the perturbation component; that is, new pairs of geometrically complementary active triangles are sought, but using a smaller d value. Our recent work on docking shows, however, that it becomes difficult to find geometrically complementary active triangles with smaller values of d [72]. A more effective alternative is to sample new rigid-body transformations directly rather than through new pairs of geometrically complementary active triangles and to do so in a continuous small neighborhood of an initial transformation.

Let the vector $\langle t, u, \theta \rangle$ be a rigid-body transformation, where t refers to the translation component, and $\langle u, \theta \rangle$ is an axis-angle representation of the orientation component (implemented through quaternions). In each move in the local optimization, a new random transformation is sampled in a small neighborhood of the transformation representing the configuration resulting from the previous modification. A new translation component t' is sampled in a δ_t neighborhood of t. A new axis u' is sampled by rotating around u by a sampled angle value δ_ϕ; a new angle θ' for the rotation component is obtained by sampling in a δ_θ neighborhood around θ. The result is that each move is a small modification of the contact interface to project a configuration onto its nearest local minimum. We note that, as before in the context of structure prediction, a working definition is employed here for the local minimum.

The result is a trajectory of low-energy dimeric configurations that are useful as decoys for the purpose of docking protocols. As in protein structure prediction, protein-protein docking protocols rely on first obtaining low-energy decoy configurations. Structural and energetic analysis then allows selecting a subset for further refinement in order to make a prediction on the native quaternary structure.

2.3. BH for Mapping Minima Connecting Diverse Stable States of a Protein Molecule. Many proteins employ motions to access different structures that allow them to tune their biological function [73, 74]. An important problem is to understand how a protein transitions between different functionally relevant states [50, 75, 76]. The problem of obtaining transition trajectories is directly related to that of obtaining the connectivity of the space around stable states. Computing transition trajectories is challenging [77], as such trajectories can connect structural states far away in the variable space. By taking into account a system's dynamics, the typical MD framework is in principle desirable to provide information on the time scales associated with conformational changes in a transition trajectory. However, its practical application is limited. Long simulation times may be needed to observe a transition trajectory to go over energy barriers.

In this proof-of-concept application, we propose that the BH framework can be a valuable tool as a first step towards elucidating transition trajectories. BH can be employed to map the minima connecting two given structural states and thus elucidate energetically credible conformational paths. Treating conformations in the path as important milestones, MD-based techniques can then be employed to locally deform a conformational path into an actual transition trajectory that incorporates dynamics [78].

In this application, the representation, energy function, and the implementations of the perturbation and local search

components of BH are as in Section 2.1. Here we pursue a proof-of-principle demonstration as follows. Let us suppose we are given two stable structural states of a protein, A and B. One of them can be regarded as the initial conformation to initiate a BH trajectory of local minima, and the other as the goal. A given number, let us say h, of BH trajectories can be initiated from the initial structure. The trajectories are allowed to grow for a fixed number of energy evaluations. In the unbiased scenario, the trajectories do not employ information about the location of the goal conformation in variable space. The results in Section 3 show that with sufficient sampling, if the initial and goal conformations are low-energy (i.e., stable), even unbiased BH trajectories are successful at approaching the goal conformation.

In a second scenario, the trajectories can be biased. Let us define an ϵ-radius ball around the goal conformation. As long as a BH trajectory stays outside this volume of the variable space (i.e., no minima are ϵ or closer to the goal), the BH exploration proceeds unbiased. When the trajectory enters the designated goal region of the variable space, say through its current minimum C_i, the exploration is biased towards obtaining minima that stay within the goal region. Given C_i, multiple perturbations followed by local optimization are attempted until a C_{i+1} is found which remains in the goal region. While the number of attempts is limited to a maximum of l consecutive failures before the BH exploration returns to its unbiased setting, in practice it is possible to remain in a goal region for a sufficiently large ϵ. The exploration terminates when the goal conformation is approached within some determined tolerance.

The value of ϵ is related to that of l. Moreover, a meaningful value for ϵ depends on the distance metric used and its effectiveness on a particular system. For instance, on small proteins, lRMSD can be used to determine the radius of the goal region. On other systems, instead, other measurements allow circumventing some of the issues with lRMSD. For instance, the TM-score [79] and GDT_TS [80] allow better capturing structural similarities than lRMSD when motions are localized to specific regions. The two are also less sensitive to noise. Familiarity with the system to be modeled allows better determining which measurement should be used and what values will be effective for ϵ and l. As the goal in this paper is to show a proof-of-concept demonstration that BH can be useful to obtain information on minima connecting diverse stable states in the equilibrium conformational space of a protein, we do not devote time to fine tuning parameters. The results in Section 3 show that values exist for these parameters that allow BH to come in closer proximity to the goal structural state in the biased over the unbiased implementation. Further tuning of the parameters is expected to improve the results and provide interesting directions for researchers to explore the viability of BH in this application context.

3. Results

Experimental Setup. The stopping criterion in each experimental setting to evaluate the performance of BH is set to a fixed number of energy evaluations. This number is 10^7 energy evaluations for the application of BH on structure prediction and 10^6 on our last application of connecting between different stable states. Results for the protein-protein docking do not change after 10,000 conformations, so this number is employed as a stopping criterion. Additionally, m and k are set to 100 and 20, respectively. In the application on docking, different values are tried for δt, $\delta\phi$, $\delta\theta$, and the ones employed for the experiments presented below are 1.5 Å, 10°, and 30°, respectively. On the last application of BH, h is set to 10 trajectories, ϵ is set to a TM-score of 0.4, $l = 100$, and the exploration terminates earlier than 10^6 energy evaluations when the current minimum is within a TM-score of 0.9 of the goal conformation.

We present three main sets of results according to the three different BH applications analyzed here. Where possible, results are compared to those reported by other state-of-the-art structure prediction or protein-protein docking methods (Tables 1 and 2 in the results presented in Sections 3.1 and 3.2, resp.). In addition, analysis of the BH-obtained minima is conducted, and distributions of the distances between consecutive minima are shown. This allows evaluating whether the implementations for the perturbation and local optimization in each application setting preserve the adjacency relationship between consecutively obtained minima. The comparison with state-of-the-art methods and the adjacency analysis employ the least Root-Mean-Squared Deviation (lRMSD) semimetric. Analysis of results obtained on the third application of BH on connecting stable states of a protein molecule employs additional measurements, such as GDT_TS and TM-score (Tables 3 and 4 in the results presented in Section 3.3). The minima sampled by BH in the context of this third application are also visualized on a low-dimensional projection of the variable space (the projection coordinates are detailed below) that reveals where the BH sampling focuses.

The main measurement used in the analysis below is lRMSD. Briefly, lRMSD measures the weighted Euclidean distance between corresponding atoms after optimal superposition of the two conformations under comparison (or configurations, if consisting of more than one polypeptide chain). The optimal superposition refers to the rigid-body motion or transformation in SE(3) minimizing this weighted Euclidean distance [81]. lRMSD captures structural dissimilarity, but it is not a Euclidean metric, as it does not obey the triangle inequality. Low values indicate high similarity, and high values indicate high dissimilarity, but interpretation of intermediate values is difficult. Interpretation has been the subject of many studies [82]. For instance, lRMSD has been found to depend on system size. A 5 Å lRMSD between a computed conformation and the native structure of a short protein chain of no more than 30 amino acids is considered a large deviation, but the same dissimilarity is less significant for a medium-size protein of 100 amino acids or more. Working interpretations abound. In general, for medium-size proteins, if the lowest lRMSD obtained over computed conformations to the known native structure is more than 6-7 Å, the native structure is not considered to have been captured in silico.

Table 1: Comparison of the lowest lRMSDs obtained by BH to those obtained by other methods on the protein dimers studied here. MBH refers to monotonic BH. lRMSDs reported by BH, MBH, and the work in [56] in columns 5–7 are over backbone atoms, whereas those reported by the work in [36, 84] in columns 8-9 are over alpha carbons of the backbone chain.

Number	PDB ID	Length	Fold	BH (Å)	MBH (Å)	[56] (Å)	[36] (Å)	[84] (Å)
1	1dtdB	61	α/β	6.9	6.6	7.5	6.5	5.7
2	1isuA	62	α/β	6.3	6.5	6.5	6.5	6.9
3	1c8cA	64	α/β	6.5	5.7	7.2	3.7	5.0
4	1sap	66	α/β	6.5	6.0	7.36	4.6	6.6
5	1hz6A	67	α/β	5.7	6.0	6.6	3.8	3.4
6	1wapA	68	β	7.4	8.1	7.3	8.0	7.7
7	1fwp	69	α/β	6.3	6.7	7.1	8.1	7.3
8	1ail	70	α	3.2	4.2	4.0	5.4	6.0
9	1aoy	78	α/β	5.7	6.1	5.8	5.7	5.7
10	1cc5	83	α	5.8	5.6	5.8	6.5	6.2
11	2ezk	93	α	4.3	5.8	6.0	5.5	6.6
12	1hhp	99	β	10.4	10.5	11.0	NA	NA
13	2hg6	106	α/β	8.8	9.3	9.7	NA	NA
14	3gwl	106	α	4.9	4.9	6.3	NA	NA
15	2h5nD	123	α	7.5	7.8	8.6	NA	NA

Table 2: Comparison of the lowest lRMSDs obtained by BH to those obtained by other methods. Systems that are CAPRI targets are denoted by an asterisk.

Number	PDB ID (chains)	Size	BH (Å)	[66] (Å)	[91] (Å)
1	1c1y (A,B)	1376, 658	1.8	1.2	N/A
2	1ds6 (A,B)	1413, 1426	3.4	1.2	N/A
3	1tx4 (A,B)	1579, 1378	2.4	1.4	N/A
4	1www (W,Y)	862, 782	2.6	11.4	N/A
5	1flt (V,Y)	770, 758	2.7	1.5	N/A
6	1vcb (A,B)	755, 692	3.4	0.8	N/A
7	1vcb (B,C)	692, 1154	2.7	13.1	N/A
8	1ohz* (A,B)	1027, 416	2.7	1.7	0.6
9	1t6g* (A,C)	2628, 1394	3.6	1.7	3.8
10	1zhi* (A,B)	1597, 1036	4.6	25.3	3.4
11	2hqs* (A,C)	3127, 856	2.6	29.1	2.5
12	1qav (A,B)	663, 840	2.6	1.4	N/A
13	1g4y (B,R)	682, 1156	4.1	0.8	N/A
14	1cse (E,I)	1920, 522	2.4	0.7	N/A
15	1g4u (R,S)	1398, 2790	3.2	1.0	N/A

However, high values of lRMSD cannot be automatically interpreted to indicate significant structural dissimilarity. Since lRMSD weighs each atom equally, it cannot capture global topology changes and overly penalizes cases where the differences are localized to a specific region of the molecule due to, say, a large-scale motion. For instance, the lRMSD of two conformations can be high even if structural deviations are limited to a loop that has a different orientation in the two conformations under comparison [83]. In such cases, other measurements, such as GDT_TS and TM-score, can be more appropriate. While different in implementation details, these two scores essentially locate a maximum subset of atoms between two conformations under comparison which are close in space after optimal superposition and minimizes an overall lRMSD-based error. While GDT_TS is reported in %, TM-score is unitless. Both capture similarity, so higher values are better. While lRMSD and GDT_TS depend on system size, TM-scores are found to be more reliable [83], which is why we employ them here in the analysis in Section 3.3 on the third BH application on connecting diverse stable states.

3.1. Analysis of BH-Obtained Decoy Conformations of a Protein Polypeptide Chain. Our realization of the BH framework for the purpose of ab-initio structure prediction is applied to a comprehensive list of 15 target protein systems. These systems, listed in Table 1, range from 61–123 amino acids in length and cover the α, β, and α/β folds. Many of them are selected due to the availability of data reported on them by structure prediction protocols. On these systems, computing 10^7 energy function evaluations takes 1–4 days of CPU time on a 2.4 Ghz Core i7 processor, depending on chain length.

3.1.1. Comparison with State-of-the-Art Methods. Table 1 shows comparisons to state-of-the-art methods in ab-initio structure prediction in terms of lRMSD. Over all minima obtained from each amino acid sequence by BH, the conformation with the lowest lRMSD to the known native structure of that sequence (experimentally obtained structure with PDB ID shown in this table) is recorded, and that value is reported in column 5 in Table 1. To take into account stochasticity, we report in Table 1 the average lowest lRMSD obtained over 3 independent runs. This value is compared to lowest lRMSDS reported by methods that are popular in ab-initio structure prediction [36, 84]. We also compare to data obtained with our previous work on ab-initio structure prediction with a robotics-inspired tree-based exploration

Table 3: Initial conformations of calmodulin are in the rows, whereas goal conformations are denoted in the columns. Three measurements, lowest lRMSD, highest TM-score, and highest GDT_TS scores to any of the goal conformations, are reported over the 10 trajectories started from each initial conformation.

PDB	1cfd			1cll			2f3y		
ID	lRMSD (Å)	TM-score	GDT_TS (%)	lRMSD (Å)	TM-score	GDT_TS (%)	lRMSD (Å)	TM-score	GDT_TS (%)
1cfd	6.70	0.57	56	6.48	0.60	50	8.20	0.44	38
1cll	5.84	0.50	43	2.38	0.84	78	6.24	0.53	54
2f3y	6.7	0.47	40	2.50	0.82	74	2.87	0.77	73

Table 4: Initial conformations of adenylate kinase are in the rows, whereas goal conformations are denoted in the columns. Three measurements, lowest lRMSD, highest TM-score, and highest GDT_TS scores to any of the goal conformations, are reported over the 10 trajectories started from each initial conformation.

PDB	1dvr		2aky		2ak3		4ake	
ID	TM-score	GDT_TS (%)	TM-score	GDT_TS (%)	TM-score	GDT_TS (%)	TM-score	GDT_TS (%)
1dvr	0.99	99	0.83	74	0.41	25	0.32	20
2aky	0.84	76	1.00	100	0.44	28	0.31	18
2ak3	0.41	25	0.44	28	0.99	99	0.39	25
4ake	0.31	18	0.29	17	0.41	26	1.00	100

method [40, 41]. Monotonic BH (MBH) is also included in the comparisons (column 6).

The results in Table 1 make the case that BH performs just as well as state-of-the-art methods in structure prediction in terms of its ability to obtain low-lRMSD conformations in an ab-initio setting. The role of the energy function may partially explain some differences among the methods, as they employ different energy functions (MBH and [41] also employ AMW). It is interesting to note that, while our realization of BH (which uses a Metropolis criterion to add the next minimum to its trajectory) obtains lower lowest lRMSDs on more proteins than MBH, the performance of MBH is comparable to the other methods in many cases. MBH can be regarded as a special case of the BH framework with the Metropolis criterion, where temperature $T = 0$. The T value we use here for our realization of BH allows a 2.6 kcal/mol energy increase between two consecutive local minima with probability 0.1.

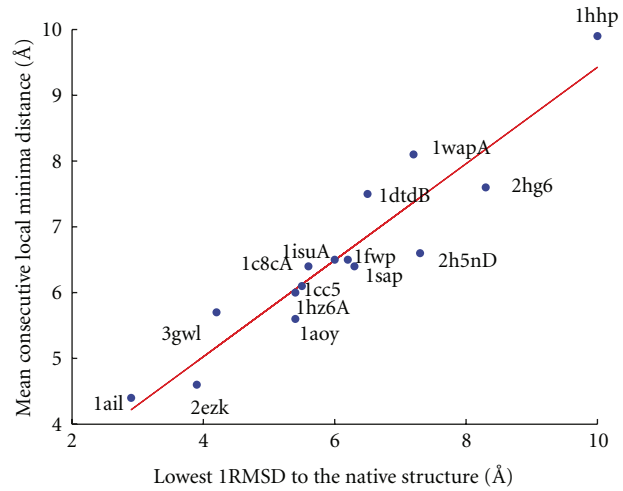


Figure 2: The mean consecutive local minima distance is drawn against the lowest lRMSD obtained for each protein.

3.1.2. Evaluation of Adjacency Relationship. Adjacency between local minima obtained consecutively by BH is often stated as important for global optimization. Here we show concretely, in the context of ab-initio structure prediction, how this adjacency correlates with the lowest lRMSD reported by BH to the known native structure of each of the protein systems studied. The lRMSD between two consecutive local minima is computed, and the average is recorded for a given protein system. This value is plotted against the lowest lRMSD from the native structure obtained by BH on each protein system in Figure 2. A strong correlation of 94% is observed in Figure 2 between the average consecutive local minima distance and the lowest lRMSD to the native structure. This result suggests that adjacency of consecutively sampled local minima is related to the ability of BH to locate the native structure. Figure 2 shows that, in cases where the average consecutive local minima distance is large, BH does not come close to the native structure.

Further detail is provided in Figure 3 on two protein systems. These systems are selected to represent two diametrical cases that correspond to the bottom left and top right portions in Figure 2. The lowest lRMSD structures obtained for each of these two systems by BH are superimposed over their respective native structures in Figures 3(a)-3(b). The entire distribution of consecutive local minima distances is shown for these two proteins in Figures 3(c)-3(d). Figures 3(c)-3(d) further show that, in cases where the majority of consecutive minima are not adjacent in variable space, the overall performance of BH in terms of lowest lRMSD to the native structure suffers. One reason for the poor adjacency is that the fragment replacement may perturb too much

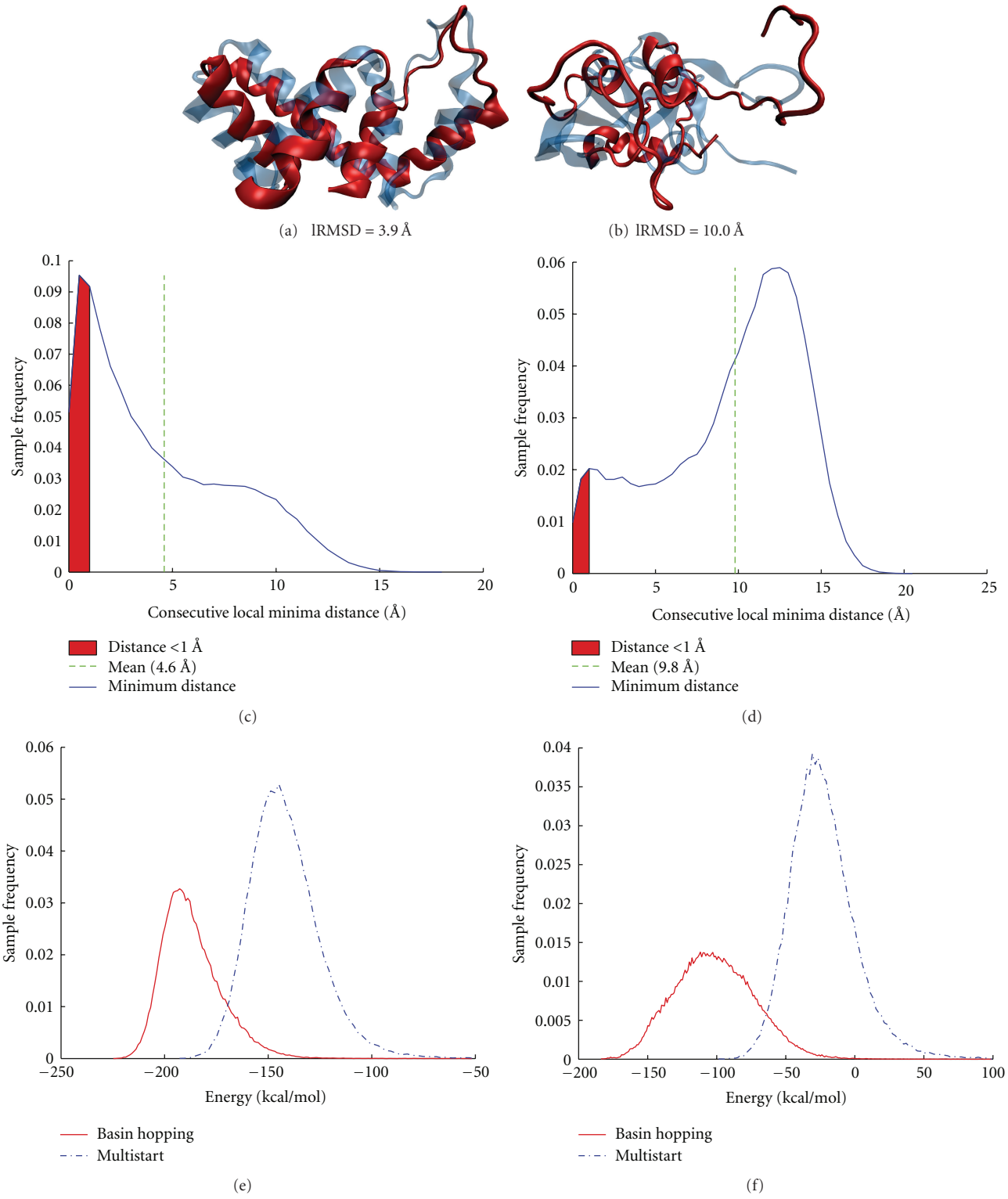


Figure 3: (a and b) The lowest-lRMSD conformation (in opaque red) is superimposed over the known native structure (in transparent blue) for the protein with native PDB ID 2ezk in (a) and 1hhp in (b). (c and d) The distribution of consecutive local minima distances in terms of lRMSD is shown in (c and d) for the two proteins, respectively. (e and f) The distribution of energies obtained by BH is superimposed over that obtained by the multistart method on each of the two proteins.

of a conformation. In a recent analysis [85], we show that this can be controlled by biasing the sampling of fragment replacements towards those that will result in small structural changes in terms of lRMSD between C_i and $C_{\text{perturb},i}$.

Comparison of BH with Multistart Sampling. Adjacency of consecutive local minima in BH is often stated as a key distinguishing characteristic over a multistart method, where initial points for local optimization are essentially sampled at random over the variable space. Here we show the effect of the adjacency relationship in a concrete setting in terms of the energetic quality of the sampled minima. On the same two protein systems where the above analysis highlights consecutive local minima distances, we show in Figures 3(e)-3(f) the distribution of energies. Figures 3(e)-3(f) superimposes the distribution obtained by BH over that obtained by the multistart method. The results show that BH obtains lower-energy minima than the multistart method. In the context of ab-initio structure prediction, the quality of decoy conformations obtained by BH is superior over that obtained by a multistart method.

Sampling Redundancy in BH. It is interesting to determine how often our realization of the BH framework here comes close to the native structure. We show this visually through a projection of the variable space in a few dimensions. The projection coordinates we choose are based on the ultrafast shape recognition (USR) features [86], which we have employed in previous work to guide a tree-based exploration of the variable space with measurements taken over a low-dimensional projection [40, 56]. These coordinates give a coarse representation of the molecular shape. They are first momenta of distance distributions of atoms in a molecule from selected points on the molecule. The selected (reference) points are the centroid (ctd), point closest to centroid (cst), point farthest from centroid (cfd), point farthest from cfd, and so on. More reference points can be defined this way, but the ones we employ for the visual representation here use only ctd and cfd. It is worth noting that, while coarse, the USR-based projection is fast to compute for each conformation, unlike PCA- or ISOMAP-based decompositions [87–89], which are time consuming and hard to use in an online setting and contain several other shortcomings noted for conformational space [90].

Figure 4 shows the projection of BH-sampled minima over the two USR-based coordinates measured using the ctd and cfd reference points. The projection is discretized so that cells can be defined in a 2d grid for the purpose of measuring how often BH samples similar minima in terms of their coarse 2d USR-based representation. The 2d grid in Figure 4 is color coded with a blue-to-red color scheme that corresponds to cells with low-to-high number of minima projected to them. The cell that contains the projection of the native structure is marked with an $\times$.

The projection of the sampled minima in this USR-based 2-dimensional space allows visualizing highly sampled regions by BH. The representation is coarse (e.g., cell widths used here for visualization can be made smaller), as conformations that map to the same cell may be several Å apart, but the projection is useful to draw two conclusions. First, compared to the vast variable space (sea of blue in Figure 4), BH sampling seems to focus in regions near the native structure. These regions represent the equilibrium conformational space. Second, sampling can be redundant; some regions are more populated than others. Future research can address redundancy in order to enhance the capability of BH to sample the equilibrium conformational space of a protein molecular in terms of local minima.

3.2. Analysis of BH-Obtained Decoy Configurations of Protein-Protein Dimers. Our realization of the BH framework for the purpose of protein-protein docking is applied to a comprehensive list of 15 different dimers. These vary in size, represent diverse functional classes, and have been tested by other protein-protein docking methods, and some are even CAPRI targets. Testing is carried out on a 2.66 GHz Opteron processor with 8 GB of memory. Depending on system size, obtaining 10,000 conformations takes 6–12 CPU hours.

3.2.1. Comparison with State-of-the-Art Methods. Table 2 shows the lowest lRMSD from the known native structure (with PDB ID shown in column 1) obtained by BH in column 3. Lowest lRMSDs reported on these systems by other methods are shown in columns 4-5. System size in terms of number of atoms in each of the chains is shown in column 2. Table 2 shows that BH achieves low lRMSDs to the native structure on each system. Moreover, these are comparable to the lRMSDs reported by other related methods. In particular, the method presented in [66] employs geometric hashing, whereas that in [91] uses long optimizations with a carefully designed energy function that employs information on evolutionary conservation to sample low-energy conformations. In addition to a comparable performance with these methods, BH samples many configurations within 5 Å lRMSD of the native structure (data not shown). These configurations, if selected and further refined in the course of a multistage docking protocol, will allow obtaining the native structure in great detail.

3.2.2. Evaluation of Adjacency Relationship. We investigate here the adjacency between consecutively sampled local minima. Figure 5(a) plots the mean consecutive local minima distance in terms of lRMSD for each protein against the lowest lRMSD obtained to the native structure. A positive correlation of 73% is observed. The mean consecutive local minima distance is less than 15 Å for about half of the systems. While this may seem like a large number compared to the related results on ab-initio structure prediction, the range is larger due to the size of the dimeric systems (lRMSD depends on size). The strong correlation suggests that adjacency of consecutively sampled minima directly relates with the ability of BH to locate a global minimum. Lower lowest lRMSDs (<5 Å) are obtained here compared to the ab-initio structure prediction setting. This is not surprising, as the variable space here is 6-dimensional, whereas the space

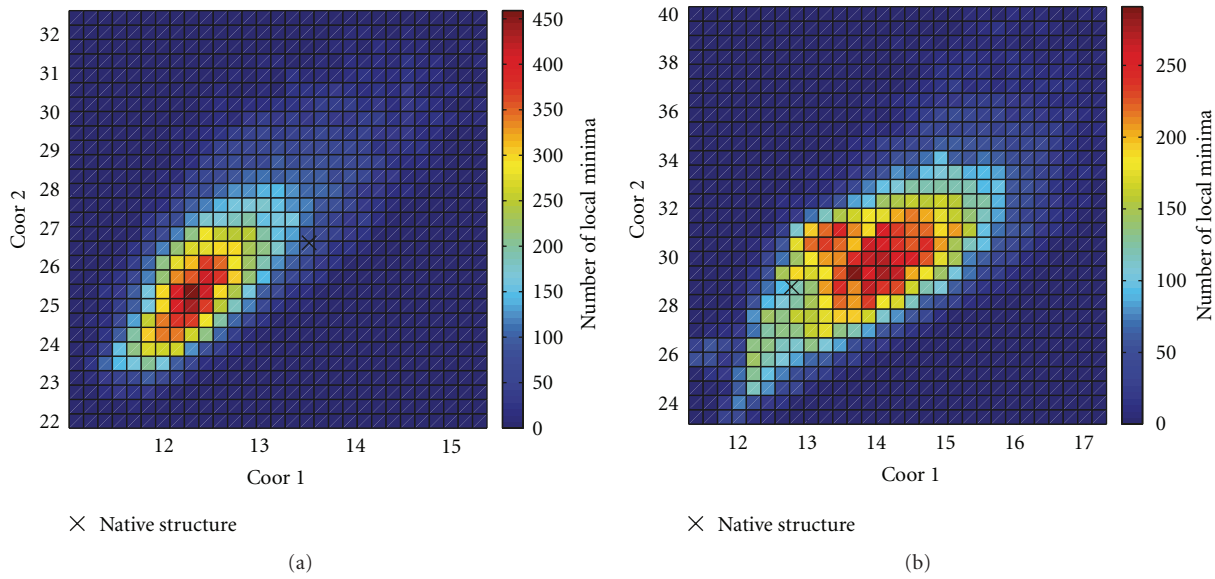


Figure 4: The 2d grids show projections of BH-sampled minima using two USR-based coordinates (with ctd and cfd as reference points). The projection for the protein with native PDB ID 2ezk is shown in (a), and that for the protein with native PDB ID 1hhp is shown in (b). The grids are color coded with a blue-to-red color scheme to show cells with low-to-high number of minima projected to them. The cell that contains the projection of the native structure is marked with an ×. The range of values of each of the coordinates is estimated as in [40]. Maximum values are based on an extended chain, and minimum values are based on the Flory compact self-excluding model of a chain of n amino acids. To improve visualization, the ranges are limited here, and the grids are clipped to allow focusing to regions with some minimal population.

in the ab-initio structure prediction application contains hundreds of dimensions.

Further detail is provided in Figures 5(d)-5(e), which shows the distribution of lRMSDs between consecutively sampled local minima on two protein systems. These systems are selected to represent two diametrical cases that correspond to the bottom left and top right portions in Figure 5(a). The actual lowest-lRMSD structures obtained on these systems are shown in Figures 5(b)-5(c), superimposed over the corresponding native structures of these proteins. The distributions in Figures 5(d)-5(e) show that more pairs of consecutive minima with low lRMSDs are obtained for the protein where BH also obtains a lower lowest lRMSD to the native structure.

3.3. Analysis of BH Trajectories in Connecting Diverse Stable States of a Protein. The unbiased setting of BH is tested here in detail on two proteins, calmodulin and adenylate kinase. Some encouraging results are shown for the biased setting as well, but a detailed investigation of the biased implementation and parameter tuning is beyond the purpose of this work.

3.3.1. Mapping Minima between Stable States in Calmodulin. Calmodulin is a 144 amino-acid long EF-hand protein that binds calcium and regulates more than 100 proteins, including kinases, phosphodiesterases, calcium pumps, and motility proteins [46–48]. The protein resembles a dumbbell, with the terminal domains linked by a flexible α-helix and the termini in a transorientation from each other on either side of the central linker. The partial unfolding of the central linker around position 77 gives calmodulin flexibility.

Calmodulin has been captured in three different functionally relevant structural states in the wet laboratory [92–94]. These states are documented in the PDB as X-ray structures under PDB IDs 1cfd (apo state), 1cll (calcium-binding state), and 2f3y (collapsed peptide-binding state). The central helix is fully formed in the calcium-binding state, unfolds in the middle in the apo state, and bends in the collapsed state. Transitions between the apo and collapsed states have been observed both in experiment and simulation [49, 50].

In order to test the unbiased setting of BH in this application, the following experiment is conducted. Each of the structures is obtained from the PDB and employed as an initial conformation. $h = 10$ Bh trajectories are launched independently from any of them. The proximity to any of the other two structures is reported in Table 3. Entry at row i and column j reports the best proximity over the 10 trajectories initiated from structure i to goal structure j.

Proximity to the goal is measured in three ways. Table 3 shows lowest lRMSD, highest TM-score [79], and highest GDT_TS [80]. The results in Table 3 show that BH is able to capture the structure with PDB ID 1cfd (apo) when initiated from 1cll (semicollapsed state) and vice versa (TM-scores above 0.5 are found to capture significant structural similarity [83]). BH also captures the structure with PDB ID 1cll (semicollapsed state) when initiated from 2f3y (collapsed state) and vice versa. A very low lRMSD and very high TM-score and GDT_TS score are obtained from 2f3y to 1cll.

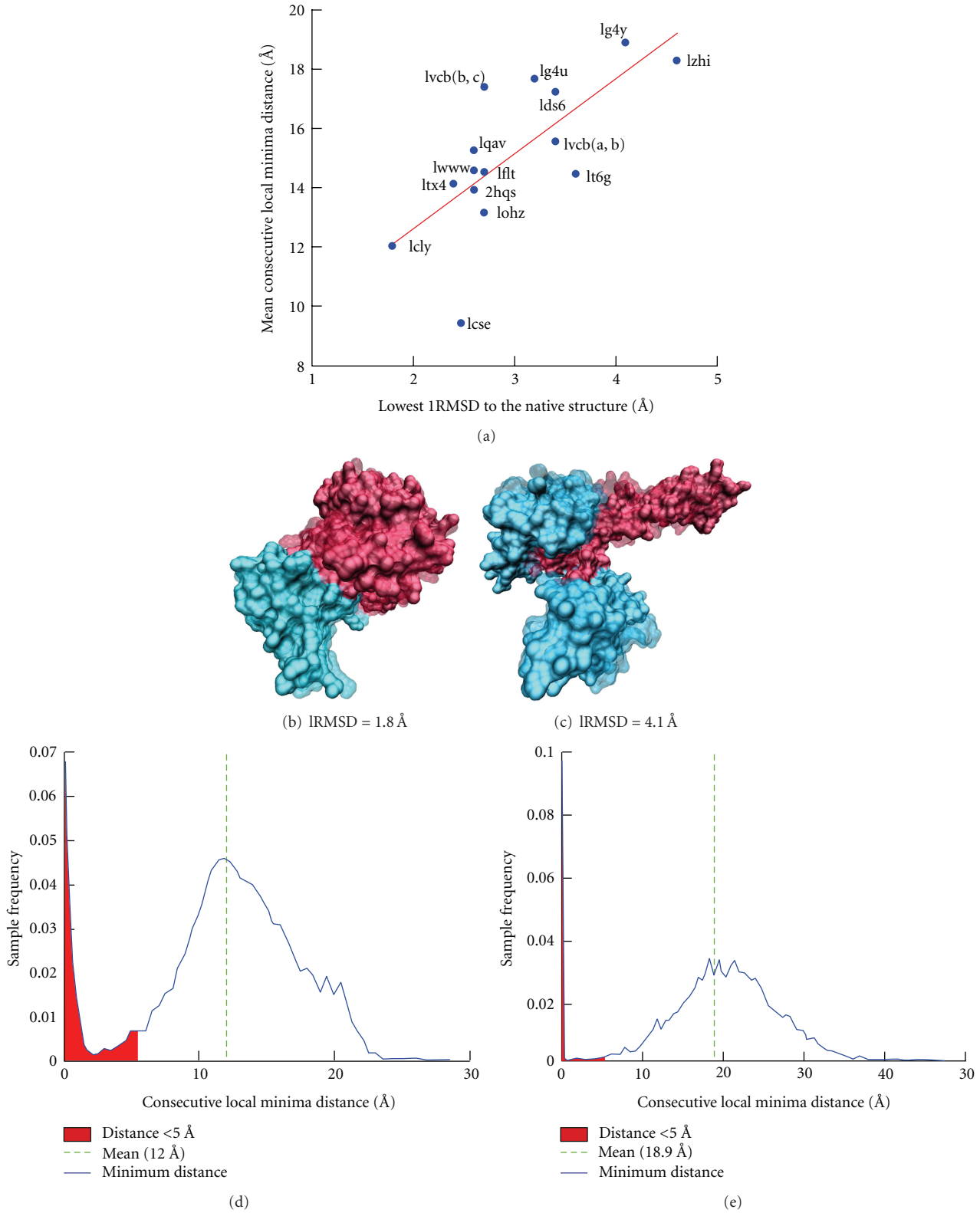


Figure 5: (a) The mean consecutive local minima distance is drawn against the lowest lRMSD obtained for each protein. (b and c) The lowest-lRMSD dimeric configuration (in opaque, with chains in different colors) is superimposed over the known native structure (in transparent) for the protein with native PDB ID 1c1y in (b) and 1g4y in (c). (d and e) The distribution of consecutive local minima distances in terms of lRMSD is shown in (d and e) for each of the proteins, respectively.

This is an encouraging result, since 1cll captures a partially closed state, whereas 2f3y captures calmodulin in its close state. Structurally, 1cfd, the apo state, is further away from the semicollapsed and collapsed states. Indeed, all three measurements in Table 3 indicate that BH has not captured state 1cfd from 2f3y and vice versa.

Table 3 also shows values along the main diagonal, which record the closest that a BH trajectory comes to its initial conformation. While this is achieved often through the very first minimum, the structure with PDB ID 1cfd is the only exception. This indicates that this structure is not at a local minimum, and BH quickly steers away from this initial conformation. This may also explain the difficulty of capturing that state when initiated from any of the other two. In addition, analysis of the biased setting for calmodulin reveals that when an ϵ value of 0.4 in terms of TM-score is used, BH is able to come closer to the goal structures. Improvements are around 0.5 Å (data not shown).

3.3.2. Mapping Minima between Stable States in Adenylate Kinase. Adenylate kinase is a 214 amino-acid long phosphotransferase enzyme that maintains energy balance in cells by catalyzing the reversible reaction $Mg^{2+} \cdot ATP + AMP \rightleftharpoons Mg^{2+} \cdot ADP + ADP$ [95]. The protein consists of a CORE domain and two substrate-binding (AMP- and ATP-) domains. The binding domains move and bind substrates independently, resulting in different functional states.

Adenylate kinase has been found in four different structural states in the wet laboratory [96–99]: the apo state, where both substrate-binding domains are open (available in the PDB under PDB ID 4ake), the collapsed state, where both domains are closed (available under PDB ID 2aky), and two intermediate states, where one of the domains is open and the other closed (PDB IDs 1dvr and 2ak3). Transitions between the apo and collapsed states have been observed both in experiment and simulation [76, 100, 101].

Unbiased BH trajectories are initiated from each of the four structures, and the best proximity to any of the other three is reported in Table 4 in terms of TM-score and GDT_TS. lRMSD is not employed, as the chains deposited under the PDB IDs listed above are of different lengths (due to differences in the setup of the structure resolution protocol in the wet laboratory). The results in Table 4 show that adenylate kinase is indeed a challenging system. The BH trajectories manage to come close only to the structure with PDB ID 2aky when initiated from the structure with PDB ID 1dvr and vice versa. This is an encouraging result, nonetheless, because the structures with PDB ID 1dvr and 2aky are structurally closer to each other.

Outstanding Challenges. Calmodulin and adenylate kinase are considered challenging systems for computational investigation due to their large size [33]. The results above support the fact that size limits sampling capability in BH. The upper bound of 10^6 energy evaluations limits BH to sample around 1,500 minima for calmodulin and 1,000 minima for adenylate kinase. Considering the small number of minima sampled, the results above are encouraging. They suggest that, with improvements to enhance the sampling capability, BH is a promising tool for mapping the equilibrium conformational space of a protein and elucidating the connectivity between different stable states.

4. Conclusion

We have shown that BH is a general, versatile framework that allows structural characterization of important biological macromolecules, such as proteins. We have selected three different applications of importance in computational structural biology on which to show the power and promise of the BH framework. Domain-specific expertise is used to implement effective perturbation and local optimization components. Important generally recognized characteristics of the BH framework, such as adjacency of local minima and its relation to the quality of the reported global minimum, are demonstrated in the applications selected in this paper.

Taken together, the results show that BH is an effective framework for structural characterization of protein systems. It is more effective than the multistart method, and the adjacency of consecutively sampled local minima is directly related with the ability of BH to come close to the global minimum. The presented results make the case that BH can be an effective tool for generating good-quality decoys for ab-initio structure prediction and protein-protein docking and a promising framework for mapping the connectivity of functionally relevant states in flexible proteins.

We note that the implementations we offer here for the key components in BH are a first step, and further tuning can result in better performance. Further analysis into different implementations is necessary to obtain a better understanding of the BH framework and its capability to enhance sampling of molecular spaces. We are currently pursuing such a comparative analysis. For instance, in recent work [85] we show that the implementation of the perturbation component employed here is sufficient to escape a current minimum and that the greedy search employed for the local optimization is just as effective but more efficient than Metropolis MC searches at low temperature [22, 102].

The application on proteins with diverse stable states serves as a proof of concept that BH can be employed to map the intermediate minima that connect stable states. The results presented here show that the framework is promising and merits further investigation in this context. The trajectory of minima obtained by BH in connecting two stable states can be considered a coarse conformational path. This path can be transformed into an actual trajectory that takes the protein through specific molecular motions from one state to another. The process is not dissimilar from how paths in robotics motion sampling are converted to actual execution trajectories with dynamics constraints [103]. The coarse transition paths can be refined through, for instance, short steered MD simulations connecting adjacent minima. Other path deformation techniques are available, and this is a direction we will explore in future research.

We believe the exposition of BH in this paper will bring more attention to this framework as a powerful global

optimization tool for biological systems. Its versatility, as we show here in the context of three different yet related applications on proteins, merits further investigation. In particular, different implementations for the main components in BH can be investigated to balance between accuracy and efficiency. Moreover, related ideas from the evolutionary computing community on population-based strategies can be employed to promote diversity of minima, as proposed in recent work on geometrical problems [20]. Related ideas from our robotics-inspired search of molecular conformational spaces [40] can be exploited to organize the BH-sampled minima, steer the exploration away from overpopulated regions in the variable space, and so enhance the sampling capability in BH.

Acknowledgments

This work is supported in part by NSF CCF no. 1016995 and NSF IIS CAREER Award no. 1144106.

References

[1] C. A. Floudas and P. M. Pardalos, *Encyclopedia of Optimization*, Kluwer Academic Publishers, Norwell, Mass, USA, 2001.

[2] "Nonconvex optimization and its applications," in *Global Optimization: Scientific and Engineering Case Studies*, J. Pinter, Ed., vol. 85 of *Mathematics and Statistics*, Springer Science and Business Media, New York, NY, USA, 2006.

[3] D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms," *Journal of Physical Chemistry A*, vol. 101, no. 28, pp. 5111–5116, 1997.

[4] A. Tiano, F. Pizzochero, and P. Venini, "A global optimization approach to nonlinear system identification," in *Conference on Control and Automation*, pp. 752–761, 1999.

[5] A. R. Leach, "A survey of methods for searching the conformational space of small and medium-sized molecules," in *Reviews in Computational Chemistry*, vol. 2, pp. 1–55, VCH Publishing, New York, NY, USA, 1991.

[6] H. A. Scheraga, "Predicting three-dimensional structures of oligopeptides," in *Reviews in Computational Chemistry*, K. B. Lipkowtiz and D. B. Boyd, Eds., vol. 3, pp. 73–142, VCH Publishing, New York, NY, USA, 1992.

[7] R. V. Pappu, R. K. Hart, and J. W. Ponder, "Analysis and application of potential energy smoothing and search methods for global optimization," *Journal of Physical Chemistry B*, vol. 102, no. 48, pp. 9725–9742, 1998.

[8] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[9] A. Nayeem, J. Vila, and H. A. Scheraga, "A comparative study of the simulated-annealing and monte carlowith-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin," *Journal of Computational Chemistry*, vol. 12, no. 5, pp. 594–605, 1991.

[10] R. C. Brower, G. Vasmatzis, M. Silverman, and C. Delisi, "Exhaustive conformational search and simulated annealing for models of lattice peptides," *Biopolymers*, vol. 33, no. 3, pp. 329–334, 1993.

[11] W. F. van Gunsteren, D. Bakowies, R. Baron et al., "Biomolecular modeling: goals, problems, perspectives," *Angewandte Chemie*, vol. 45, no. 25, pp. 4064–4092, 2006.

[12] A. Shehu, "Conformational search for the protein native state," in *Protein Structure Prediction: Method and Algorithms*, Rangwala and G. Karypis, Eds., Wiley Book Series on Bioinformatics, chapter 21, Fairfax, VA, USA, 2010.

[13] R. H. Leary, "Global optimization on funneling landscapes," *Journal of Global Optimization*, vol. 18, no. 4, pp. 367–383, 2000.

[14] M. Iwamatsu and Y. Okabe, "Basin hopping with occasional jumping," *Chemical Physics Letters*, vol. 399, no. 4–6, pp. 396–400, 2004.

[15] M. A. Miller and D. J. Wales, "Novel structural motifs in clusters of dipolar spheres: knots, links, and coils," *Journal of Physical Chemistry B*, vol. 109, no. 49, pp. 23109–23112, 2005.

[16] J. M. Carr and D. J. Wales, "Global optimization and folding pathways of selected alpha-helical proteins," *The Journal of Chemical Physics*, vol. 123, no. 23, p. 234901, 2005.

[17] T. James, D. J. Wales, and J. Hernández-Rojas, "Global minima for water clusters $(H_2O)n$, $n \leq 21$, described by a five-site empirical potential," *Chemical Physics Letters*, vol. 415, no. 4–6, pp. 302–307, 2005.

[18] R. Gehrke and K. Reuter, "Assessing the efficiency of first-principles basin-hopping sampling," *Physical Review B*, vol. 79, no. 8, Article ID 085412, 10 pages, 2009.

[19] D. J. Wales, *Energy Landscapes and Structure Prediction Using Basin-Hopping*, Wiley-VCH Verlag GmbH and Co. KGaA, 2010.

[20] A. Grosso, A. R. M. J. U. Jamali, M. Locatelli, and F. Schoen, "Solving the problem of packing equal and unequal circles in a circular container," *Journal of Global Optimization*, vol. 47, no. 1, pp. 63–81, 2010.

[21] M. Locatelli, "On the multilevel structure of global optimization problems," *Computational Optimization and Applications*, vol. 30, no. 1, pp. 5–22, 2005.

[22] B. Olson and A. Shehu, "Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface," *Proteome Science*, vol. 10, no. supplement 1, p. S5, 2012.

[23] O. M. H. R. Lourenco and T. Stutzle, "Iterated local search," in *Handbook of Metaheuristics*, F. Glover and G. Kochenberger, Eds., vol. 57, no. 513 of *Operations Research & Management Science*, pp. 321–353, Kluwer Academic Publishers, 2002.

[24] Z. Li and H. A. Scheraga, "Monte Carlo-minimization approach to the multiple-minima problem in protein folding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 19, pp. 6611–6615, 1987.

[25] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nature Structural Biology*, vol. 4, no. 1, pp. 10–19, 1997.

[26] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Current Opinion in Structural Biology*, vol. 14, no. 1, pp. 70–75, 2004.

[27] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP) round IX," *Proteins*, vol. 79, supplement 10, pp. 1–5, 2009.

[28] P. Bradley, K. M. S. Misura, and D. Baker, "Toward high-resolution de novo structure prediction for small proteins," *Science*, vol. 309, no. 5742, pp. 1868–1871, 2005.

[29] S. Yin, F. Ding, and N. V. Dokholyan, "Eris: an automated estimator of protein stability," *Nature Methods*, vol. 4, no. 6, pp. 466–467, 2007.

[30] T. Kortemme and D. Baker, "Computational design of protein-protein interactions," *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 91–97, 2004.

[31] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters," *Proteins*, vol. 65, no. 3, pp. 712–725, 2006.

[32] A. Verma, A. Schug, K. H. Lee, and W. Wenzel, "Basin hopping simulations for all-atom protein folding," *The Journal of Chemical Physics*, vol. 124, no. 4, p. 044515, 2006.

[33] A. Shehu, L. E. Kavraki, and C. Clementi, "Multiscale characterization of protein conformational ensembles," *Proteins*, vol. 76, no. 4, pp. 837–851, 2009.

[34] R. Bonneau, C. E. M. Strauss, C. A. Rohl et al., "De novo prediction of three-dimensional structures for major protein families," *Journal of Molecular Biology*, vol. 322, no. 1, pp. 65–78, 2002.

[35] T. J. Brunette and O. Brock, "Guiding conformation space search with an all-atom energy potential," *Proteins*, vol. 73, no. 4, pp. 958–972, 2008.

[36] J. DeBartolo, A. Colubri, A. K. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick, "Mimicking the folding pathway to improve homology-free protein structure prediction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 10, pp. 3734–3739, 2009.

[37] J. Debartolo, G. Hocky, M. Wilde, J. Xu, K. F. Freed, and T. R. Sosnick, "Protein structure prediction enhanced with evolutionary diversity: SPEED," *Protein Science*, vol. 19, no. 3, pp. 520–534, 2010.

[38] A. Shehu, L. E. Kavraki, and C. Clementi, "Unfolding the fold of cyclic cysteine-rich peptides," *Protein Science*, vol. 17, no. 3, pp. 482–493, 2008.

[39] M. C. Prentiss, C. Hardin, M. P. Eastwood, C. Zong, and P. G. Wolynes, "Protein structure prediction: the next generation," *Journal of Chemical Theory and Computation*, vol. 2, no. 3, pp. 705–716, 2006.

[40] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an Ab-initio tree-based exploration," *International Journal of Robotics Research*, vol. 29, no. 8, pp. 1106–1127, 2010.

[41] B. Olson, K. Molloy, and A. Shehu, "In search of the protein native state with a probabilistic sampling approach," *Journal of Bioinformatics and Computational Biology*, vol. 9, no. 3, pp. 383–398, 2011.

[42] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[43] R. Abagyan and M. Totrov, "Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins," *Journal of Molecular Biology*, vol. 235, no. 3, pp. 983–1002, 1994.

[44] P. N. Mortenson, D. A. Evans, and D. J. Wales, "Energy landscapes of model polyalanines," *Journal of Chemical Physics*, vol. 117, no. 3, pp. 1363–1376, 2002.

[45] M. C. Prentiss, D. J. Wales, and P. G. Wolynes, "Protein structure prediction using basin-hopping," *The Journal of Chemical Physics*, vol. 128, no. 22, Article ID 225106, 9 pages, 2008.

[46] A. S. Manalan and C. B. Klee, "Calmodulin," *Advances in Cyclic Nucleotide and Protein Phosphorylation Research*, vol. 18, pp. 227–278, 1984.

[47] A. R. Means, "Molecular mechanisms of action of calmodulin," *Recent Progress in Hormone Research*, vol. 44, pp. 223–262, 1988.

[48] K. T. O'Neil and W. F. DeGrado, "How calmodulin binds its targets: sequence independent recognition of amphiphilic α-helices," *Trends in Biochemical Sciences*, vol. 15, no. 2, pp. 59–64, 1990.

[49] B. E. Finn, J. Evenas, T. Drakenberg, J. P. Waltho, E. Thulin, and S. Forsen, "Calcium-induced structural changes and domain autonomy in calmodulin," *Nature Structural Biology*, vol. 2, no. 9, pp. 777–783, 1995.

[50] B. W. Zhang, D. Jasnow, and D. M. Zuckermann, "Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 46, pp. 18043–18048, 2007.

[51] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu, "Guiding protein docking with geometric and evolutionary information," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 3, Article ID 1242008, 16 pages, 2012.

[52] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr., "A graph-theory algorithm for rapid protein side-chain prediction," *Protein Science*, vol. 12, no. 9, pp. 2001–2014, 2003.

[53] M. Zhang and L. E. Kavraki, "A new method for fast and accurate derivation of molecular conformations," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 1, pp. 64–70, 2002.

[54] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Water in protein structure prediction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 10, pp. 3352–3357, 2004.

[55] A. Shehu, "An ab-initio tree-based exploration to enhance sampling of low-energy protein conformations," in *Robotics: Science and Systems*, pp. 241–248, Seattle, Wash, USA, 2009.

[56] B. S. Olson, K. Molloy, S. F. Hendi, and A. Shehu, "Guiding search in the protein conformational space with structural profiles," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 3, Article ID 1242005, 2012.

[57] J. A. Hegler, J. Laetzer, A. Shehu, C. Clementi, and P. G. Wolynes, "Restriction vs. guidance: fragment assembly and associative memory hamiltonians for protein structure prediction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 36, pp. 15302–15307, 2009.

[58] D. A. Case, T. A. Darden, T. E. I. Cheatham et al., *Amber 9*, University of California, San Francisco, Calif, USA, 2006.

[59] H. Gong, P. J. Fleming, and G. D. Rose, "Building native protein conformations from highly approximate backbone torsion angles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 45, pp. 16227–16232, 2005.

[60] K. F. Han and D. Baker, "Global properties of the mapping between local amino acid sequence and local structure in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 12, pp. 5814–5818, 1996.

[61] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nature Structural Biology*, vol. 10, no. 12, p. 980, 2003.

[62] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu, "Protein docking with information on evolutionary conserved

interfaces," in *Bioinformatics and Biomedicine Workshops (BIBMW '11)*, pp. 358–365, November 2011.

[63] G. Terashi, M. Takeda-Shitaka, K. Kanou, M. Iwadate, D. Takaya, and H. Umeyama, "The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation," *Proteins*, vol. 69, no. 4, pp. 866–872, 2007.

[64] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "PatchDock and SymmDock: servers for rigid and symmetric docking," *Nucleic Acids Research*, vol. 33, no. 2, pp. W363–W367, 2005.

[65] Y. Inbar, H. Benyamini, R. Nussinov, and H. J. Wolfson, "Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies," *Physical Biology*, vol. 2, no. 4, pp. S156–S165, 2005.

[66] Y. Inbar, H. Benyamini, R. Nussinov, H. J. Wolfson, and B. Honig, "Prediction of multimolecular assemblies by multiple docking," *Journal of Molecular Biology*, vol. 349, no. 2, pp. 435–447, 2005.

[67] S. Engelen, L. A. Trojan, S. Sacquin-Mora, R. Lavery, and A. Carbone, "Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling," *PLoS Computational Biology*, vol. 5, no. 1, Article ID e1000267, 2009.

[68] M. L. Connolly, "Analytical molecular surface calculation," *Applied Crystallography*, vol. 16, no. 5, pp. 548–558, 1983.

[69] R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov, "Examination of shape complementarity in docking of unbound proteins," *Proteins*, vol. 36, no. 3, pp. 307–317, 1999.

[70] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: a program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[71] T. Kortemme and D. Baker, "A simple physical model for binding energy hot spots in protein-protein complexes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 14116–14121, 2002.

[72] I. Hashmi and A. Shehu, "A basin hopping algorithm for protein-protein docking," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM '12)*, J. Gao, W. Dubitzky, C. Wu et al., Eds., pp. 466–469, Philadelphia, Pa, USA, 2012.

[73] J. R. Schnell, H. J. Dyson, and P. E. Wright, "Structure, dynamics, and catalytic function of dihydrofolate reductase," *Annual Review of Biophysics and Biomolecular Structure*, vol. 33, pp. 119–140, 2004.

[74] E. Z. Eisenmesser, O. Millet, W. Labeikovsky et al., "Intrinsic dynamics of an enzyme underlies catalysis," *Nature*, vol. 438, no. 7064, pp. 117–121, 2005.

[75] K. I. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes, "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: structure-based molecular dynamics simulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 32, pp. 11844–11849, 2006.

[76] Q. Lu and J. Wang, "Single molecule conformational dynamics of adenylate kinase: energy landscape, structural correlations, and transition state ensembles," *Journal of the American Chemical Society*, vol. 130, no. 14, pp. 4772–4783, 2008.

[77] P. Majek, H. Weinstein, and R. Elber, *Pathways of Conformational Conformational Transitions in Proteins*, chapter 13, Taylor and Francis group, 2008.

[78] D. R. Weiss and M. Levitt, "Can morphing methods predict intermediate structures?" *Journal of Molecular Biology*, vol. 385, no. 2, pp. 665–674, 2009.

[79] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins*, vol. 57, no. 4, pp. 702–710, 2004.

[80] A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3370–3374, 2003.

[81] A. D. McLachlan, "A mathematical procedure for superimposing atomic coordinates of proteins," *Acta Crystallographica A*, vol. 26, no. 6, pp. 656–657, 1972.

[82] V. N. Maiorov and G. M. Crippen, "Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins," *Journal of Molecular Biology*, vol. 235, no. 2, pp. 625–634, 1994.

[83] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?" *Bioinformatics*, vol. 26, no. 7, pp. 889–895, 2010.

[84] J. Meiler and D. Baker, "Coupled prediction of protein secondary and tertiary structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12105–12110, 2003.

[85] B. Olson and A. Shehu, "Efficient basin hopping in the protein energy surface," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '12)*, J. Gao, W. Dubitzky, C. Wu et al., Eds., pp. 119–124, Philadelphia, Pa, USA, 2012.

[86] P. J. Ballester and W. G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes," *Journal of Computational Chemistry*, vol. 28, no. 10, pp. 1711–1723, 2007.

[87] M. L. Teodoro, G. N. Phillips, and L. E. Kavraki, "Understanding protein flexibility through dimensionality reduction," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 617–634, 2003.

[88] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9885–9890, 2006.

[89] H. Stamati, C. Clementi, and L. E. Kavraki, "Application of nonlinear dimensionality reduction to characterize the confonrmational landscape of small peptides," *Proteins*, vol. 78, no. 2, pp. 223–235, 2010.

[90] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map," *Journal of Chemical Physics*, vol. 134, no. 12, Article ID 124116, 2011.

[91] E. Kanamori, Y. Murakami, Y. Tsuchiya, D. M. Standley, H. Nakamura, and K. Kinoshita, "Docking of protein molecular surfaces with evolutionary trace analysis," *Proteins*, vol. 69, no. 4, pp. 832–838, 2007.

[92] H. Kuboniwa, N. Tjandra, S. Grzesiek, H. Ren, C. B. Klee, and A. Bax, "Solution structure of calcium-free calmodulin," *Nature Structural Biology*, vol. 2, no. 9, pp. 768–776, 1995.

[93] R. Chattopadhyaya, W. E. Meador, A. R. Means, and F. A. Quiocho, "Calmodulin structure refined at 1.7 Å resolution," *Journal of Molecular Biology*, vol. 228, no. 4, pp. 1177–1192, 1992.

[94] J. L. Fallon, D. B. Halling, S. L. Hamilton, and F. A. Quiocho, "Structure of calmodulin bound to the hydrophobic IQ domain of the cardiac Cav1.2 calcium channel," *Structure*, vol. 13, no. 12, pp. 1881–1886, 2005.

[95] D. G. Rhoads and J. M. Lowenstein, "Initial velocity and equilibrium kinetics of myokinase," *Journal of Biological Chemistry*, vol. 243, no. 14, pp. 3963–3972, 1968.

[96] C. W. Müller, G. J. Schlauderer, J. Reinstein, and G. E. Schulz, "Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding," *Structure*, vol. 4, no. 2, pp. 147–156, 1996.

[97] U. Abele and G. E. Schulz, "High-resolution structures of adenylate kinase from yeast ligated with inhibitor Ap5A, showing the pathway of phosphoryl transfer," *Protein Science*, vol. 4, no. 7, pp. 1262–1271, 1995.

[98] G. J. Schlauderer, K. Proba, and G. E. Schulz, "Structure of a mutant adenylate kinase ligated with an ATP-analogue showing domain closure over ATP," *Journal of Molecular Biology*, vol. 256, no. 2, pp. 223–227, 1996.

[99] K. Diederichs and G. E. Schulz, "The refined structure of the complex between adenylate kinase from beef heart mitochondrial matrix and its substrate AMP at 1.85 Å resolution," *Journal of Molecular Biology*, vol. 217, no. 3, pp. 541–549, 1991.

[100] J. Ådén and M. Wolf-Watz, "NMR identification of transient complexes critical to adenylate kinase catalysis," *Journal of the American Chemical Society*, vol. 129, no. 45, pp. 14003–14012, 2007.

[101] C. Snow, G. Qi, and S. Hayward, "Essential dynamics sampling study of adenylate kinase: comparison to citrate synthase and implication for the hinge and shear mechanisms of domain motions," *Proteins*, vol. 67, no. 2, pp. 325–337, 2007.

[102] B. Olson and A. Shehu, "Populating local minima in the protein conformational space," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM '11)*, pp. 114–117, November 2011.

[103] H. Choset, K. M. Lynch, S. Hutchinson et al., *Principles of Robot Motion: Theory, Algorithms, and Implementations*, MIT Press, Cambridge, Mass, USA, 1st edition, 2005.

Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2012, Article ID 585072, 12 pages
doi:10.1155/2012/585072

Research Article

CardioSmart365: Artificial Intelligence in the Service of Cardiologic Patients

Efrosini Sourla,[1] Spyros Sioutas,[2] Vasileios Syrimpeis,[3] Athanasios Tsakalidis,[1] and Giannis Tzimas[4]

[1] *Computer Engineering & Informatics Department, University of Patras, 26500 Patras, Greece*
[2] *Department of Informatics, Ionian University, 49100 Corfu, Greece*
[3] *General Hospital of Patras "Agios Andreas", 26335 Patras, Greece*
[4] *Department of Applied Informatics in Management & Economy, Faculty of Management and Economics, Technological Educational Institute of Messolonghi, 30200 Messolonghi, Greece*

Correspondence should be addressed to Efrosini Sourla, sourla@ceid.upatras.gr

Received 29 June 2012; Accepted 11 September 2012

Academic Editor: Panayiotis Vlamos

Artificial intelligence has significantly contributed in the evolution of medical informatics and biomedicine, providing a variety of tools available to be exploited, from rule-based expert systems and fuzzy logic to neural networks and genetic algorithms. Moreover, familiarizing people with smartphones and the constantly growing use of medical-related mobile applications enables complete and systematic monitoring of a series of chronic diseases both by health professionals and patients. In this work, we propose an integrated system for monitoring and early notification for patients suffering from heart diseases. CardioSmart365 consists of web applications, smartphone native applications, decision support systems, and web services that allow interaction and communication among end users: cardiologists, patients, and general doctors. The key features of the proposed solution are (a) recording and management of patients' measurements of vital signs performed at home on regular basis (blood pressure, blood glucose, oxygen saturation, weight, and height), (b) management of patients' EMRs, (c) cardiologic patient modules for the most common heart diseases, (d) decision support systems based on fuzzy logic, (e) integrated message management module for optimal communication between end users and instant notifications, and (f) interconnection to Microsoft HealthVault platform. CardioSmart365 contributes to the effort for optimal patient monitoring at home and early response in cases of emergency.

1. Introduction

Internet has broaden the scope of medical information systems and led to the development of distributed and interoperable information sources and services. In the same time, the need for standards became crucial. Federated medical libraries, biomedical knowledge bases, and global healthcare systems offer a rich information sink and facilitate mobility of patients and practitioners [1].

Medical information systems (MISs) may include medical imaging storage and transmission systems, nursing information systems, laboratory information systems, and pharmacy information systems. To treat patients, medical personnel can use different information systems in accordance with their needs, in order to diagnose and run tests, like blood tests, urine sampling, computed tomography scans, X-ray [2], and so on. A medical information system produces all kinds of medical information in various formats, including texts, numbers, pictures, and static and dynamic images. This heterogeneous information can then be integrated without the need of medical personnel. According to a patient ID, name, or other basic data, the information can be indexed by, for medical use upon request [3].

Additionally, the spectacular penetration of mobile phones in the technological arena and their transformation into smartphones have introduced a new field of software applications' development. Smartphones have been employed widely in health care practice [4]. The level of their use is expected to increase, especially if they are enriched with doctor suitable functions and software applications. The lack of such applications is noticed even in countries with

leadership role in mobile technologies, as it is mentioned in [5]. The impact of mobile-handheld technology on hospital physicians' work practices, and patient care is systematically reviewed in [6], where the authors recommend future research about the impact of the mobility devices on work practices and outcomes.

An example which successfully combines MISs with the advantages and capabilities of Smartphones, in Orthopedics, is the integrated system that was developed for recording, monitoring, and studying patients with open tibia fractures [7]. The authors participated in the development of the system, which is based on web and mobile applications. Primary goal was the creation of a system that contains most of the scientifically validated data elements, reducing in this way omission and improving consistency, by standardizing the reporting language among medical doctors. The system's web and mobile interfaces are designed to require almost no text entry and editing and are based on the traditional medical way of acting, thus making it a doctor friendly system.

Artificial intelligence (AI) has significantly contributed in the evolution of medical informatics and biomedicine since it provides a variety of tools available to be exploited, from rule-based expert systems and fuzzy logic to neural networks and genetic algorithms. The earliest work in medical AI dates back to the early 1970s, when the field of AI was about 15 years old. Since then, there is a growing interest in the application of AI techniques in biomedical engineering and informatics, ranging from knowledge-based reasoning for disease classification to learning and discovering novel biomedical knowledge for disease treatment, indicative of the maturity and influence that have been achieved to date [8]. In [9], the integration of AI techniques in biomedical engineering and informatics is presented, especially in the following core topics: (a) feature selection, (b) visualization, (c) classification, (d) data warehousing and data mining, and (e) analysis of biological networks. In literature, a great number of research projects reflect the integration of AI in medicine, from the use of fuzzy expert systems [10, 11] and design patterns [12], to neural networks [4, 13] and decision support systems (DSSs) [14, 15].

This paper presents an integrated system (CardioSmart365) based on web applications, web services, and smartphone applications for lifelong cardiologic patient monitoring, early detection of emergency, and optimal process management of the emergency incident. Cardiologic patient modules with DSSs based in fuzzy logic are developed for the most common heart diseases. The system allows interaction and communication between cardiologic patients, cardiologists, general practitioners, hospitals, and outHospitals health sectors. Everyday clinical practice, medical doctors, cardiologic patients, research and science, and healthcare systems benefit from the proposed system.

The rest of the paper is organized as follows: related work is presented in Section 2 followed by motivation in Section 3. Section 4 presents literature information about the different types of health records, the existing health record providers, and their evaluation. This section can be omitted for expert users. Section 5 presents a thorough description of the implemented system, including its architecture, functionality, components, and software framework. System's added value is discussed in Section 6. Finally, future steps are proposed in Section 7, and the paper concludes in Section 8.

2. Related Work

2.1. Mobile Health Applications. Most health applications in online markets are native applications, patient oriented or medical doctor (MD) oriented. In most cases, the patient-oriented health applications are exploited only by patients, and the information gathered is not available directly to physicians, through a communication channel. Moreover, the MD-oriented health applications serve specific purposes, mostly for educational and quick access to medical literature reasons. On the other hand, frequently, mobile applications that are part of medical research projects, store information, and send it to collaborative servers for additional processing and disposal to physicians.

The use of individual mobile health applications, developed to serve specific purposes, is widely spread. The need for such applications is apparent in every major online market for mobile applications including Android Market, Apple Store, and Samsung Apps. Applications developed for cardiology record blood pressure and cardiac pulses, applications for diabetes record blood glucose [16], and for obesity, they record calories and diet [17], for dementia they use GPS to monitor the patient [18, 19], and applications for chronic diseases target mobile phones with sensors and detect tachycardia or respiratory infections [20].

2.2. Medical Applications for Cardiology. Many mobile applications for cardiology have been developed in order to enhance medical doctors' and medical students' research experience [21–23] such as (a) applications that present a 3D prototype of a human heart and allow users to observe the heart from any angle, (b) calculators with commonly used formulas in cardiovascular medicine, (c) electrocardiography (ECG) guides with samples of different types of ECG, (d) guideline tools for clinical practice and diagnosis, and (e) decision support tools including several criteria and cases. All the above applications are addressed to medical staff, mainly for educational reasons and quick access to literature data, useful for medical doctors. In addition, they are not suitable for cardiologic patients.

A web environment for monitoring cardiologic patients is Heart360 Cardiovascular Wellness Center [24], sponsored by the American Heart Association and American Stroke Association. Heart360 allows patients to monitor their blood pressure, blood glucose, cholesterol, weight, nutrition, and physical activity, while receiving education and information specific to their condition. Heart360 utilizes Microsoft HealthVault [25]. More specifically, patients are able to (a) collect and record their blood pressure, blood glucose, cholesterol, weight, nutrition, and physical activity habits, (b) set goals and track their progress, (c) view their data in charts and graphs that they can print out and share with others involved in their family health, (d) manage multiple user accounts, and (e) get news and articles of

potential interest based on their store of health information. This application is patient oriented and does not offer any substantial help to cardiologists or general practitioners.

2.3. Artificial Intelligence in Cardiology. Recently, AI, out of invasive and noninvasive diagnostic tools, becomes the promising method in the diagnosis of heart diseases. In [4], a comparison is presented of multilayered perceptron neural network (MLPNN) and support vector machine (SVM) on determination of coronary artery disease (CAD) existence upon exercise stress testing (EST) data. In [13] neural networks are used as the most suitable solution to outcome prediction tasks in postoperative cardiac patients. An AI-based Computer Aided Diagnosis system is designed in [26] to assist the clinical decision of nonspecialist staff in the analysis of heart failure patients. The system computes the patient's pathological condition and highlights possible aggravations, using four AI-based techniques: a Neural Network, a support vector machine, a decision tree and a fuzzy expert system whose rules are produced by a Genetic Algorithm. Neural networks achieved the best performance with an accuracy of 86%. Another application domain for AI is nuclear cardiology imaging, since the automatic interpretation of nuclear cardiology studies is a complex and difficult task, and a variety of expert systems, neural networks, and case-based reasoning approaches have been attempted in this area [27].

3. Motivation

Although healthcare systems have strongly benefited from the incorporation of new technologies, there is a serious lack of incorporation in the field of clinical medicine. Clinical medicine is involved with patients and their treatment, where medical doctors (MDs) are responsible for the patients' progress.

The direct implication of humans, in particular patients, presupposes that the new technologies incorporated have to be safe, reliable and to offer proven solutions. In addition, MDs are not familiar with new sophisticated applications that change their traditional way of working, justifying in this way the skepticism that MDs present in the incorporation of new technologies.

CardioSmart365 is motivated from the need to proceed in using the advantages of new technologies in the field of clinical medicine. Cardiology is a first-line emergency medical specialty that apart from the emergency incident has to deal with a variety of chronic diseases. Therefore, if CardioSmart365 is proved to be an effective tool in the hands of Cardiologists, general practitioners (GPs), and cardiology patients (CPs), then the same methodology can be used for the development of respective systems for every other medical specialty.

Moreover, to the best of the authors' knowledge, till now there is no any system available that incorporates the following characteristics: (a) to be both patient and medical doctor oriented, (b) to incorporate AI modules, (c) to provide access to end users through multiple channels (web, mobile), (d) to utilize existing and state-of-the-art medical platforms, such as Microsoft Health Vault, and (e) to be easily expandable and deployable. Thus, the detected deficiencies and the willingness to offer enhanced services to patients and medical doctors motivated us to develop a system that incorporates these characteristics into an integrated functionality.

CardioSmart365 is part of collaboration between the Computer Engineering and Informatics Department and School of Medicine of the University of Patras and the General Hospital of Patras "Ag. Andreas" and is currently in testing phase with the following involved parties: (a) nonhospital and hospital cardiologists, (b) nonhospital and hospital general practitioners and (c) cardiologic patients. CardioSmart365 will soon be publicly available at http://www.biodata.gr/cardiosmart365/.

4. Health Records

4.1. Introduction. A personal health record (PHR) is a health record where health data and information related to the care of a patient is maintained by the patient [28]. This stands in contrast with the more widely used electronic medical record (EMR) and electronic health record (EHR) which are operated by institutions (such as hospitals) and contain data entered by clinicians or billing data to support insurance claims. The NAHIT report defines the following [29].

Electronic Medical Record (EMR). An electronic record of health-related information on an individual that can be created, gathered, managed, and consulted by authorized clinicians and staff within one health care organization.

Electronic Health Record (EHR). An electronic record of health-related information on an individual that conforms to nationally recognized interoperability standards and that can be created, managed, and consulted by authorized clinicians and staff across more than one health care organization.

Personal Health Record (PHR). An electronic record of health-related information on an individual that conforms to nationally recognized interoperability standards and that can be drawn from multiple sources while being managed, shared, and controlled by the individual.

In summary, EMRs and EHRs are tools for providers while PHRs are the means to engage individuals in their health and wellbeing.

An EMR or PHR may contain a fairly-wide range of information related directly or indirectly to the health of the user [28], more specifically:

(i) personal information, that is, name, date of birth, and current address,

(ii) names and phone numbers of relatives or people of the owner's friendly environment that can be contacted in case of emergency,

(iii) names, addresses, and telephone numbers of physicians,

(iv) info related to individual health insurance,

(v) current medication (if any) and respective dosages,

(vi) known allergies to foods, drugs, and other substances,

(vii) important events, dates, and hereditary diseases, involving the family history,

(viii) history of the most important diseases encountered by the user in his past,

(ix) results of medical examinations, important medical tests, dental history as well as vaccination history,

(x) recent medical diagnostics: summary of visits to family physician or another specialist,

(xi) information related to physical activity, exercise program, dietary restrictions, and record of medications that do not require a prescription (over the counter - OTC) and/or alternative therapeutic approaches.

PHRs have many potential benefits to patients, caregivers, and institutions [28]. One of the most important PHR benefits is greater patient access to a wide array of credible health information, data, and knowledge. Patients can leverage that access to improve their health and manage their diseases. A critical benefit of PHRs is that they provide an ongoing connection between patient and physician, which changes encounters from episodic to continuous, thus substantially shortening the time to address problems that may arise.

The PHR can benefit clinicians in many ways. First, patients entering data into their health records can elect to submit the data into their clinicians' EHRs. The PHR may also become a conduit for improved sharing of medical records. Finally, asynchronous, PHR-mediated electronic communication between patients and members of their health care teams can free clinicians from the limitations of telephone and face-to-face communication or improve the efficiency of such personal contacts.

4.2. Health Record Providers and Evaluation. In their work, Sunyaev et al. [30] enumerate the existing PHR providers, based on US-oriented and Internet-based PHRs. The two most popular ones are Google Health and Microsoft HealthVault, which are independent products developed by profit-oriented companies or open scientific projects. In [30], these two PHR systems were chosen to be evaluated, due to the relative similarity of their architecture, target markets, and business models. A list of 25 end-user features was elicited, which are necessary for a successful PHR implementation. These features were classified in three categories: patient information, personal control, and additional services. Another work examined and compared the designs of Google Health API and Microsoft HealthVault API [31]. In the evaluation, seven different categories were used: libraries, documentation, authentication, security, data access, data modification, and data messages. The two platforms have advantages and disadvantages and present similarity in some characteristics while they are different in others. However, since Google decided to retire Google Health in January 1, 2012, we can assume that HealthVault will be dominant.

4.3. Microsoft HealthVault. Microsoft HealthVault [25] is a backend cloud-based platform, based on EMR systems, which provides a privacy- and security-enhanced foundation that can be used to store and transfer information between a variety of e-health care customer's applications (desktop, web, and mobile ones), hospital applications, and healthcare devices. It also offers tools to solution providers, device manufacturers, and developers, in order to build innovative new health and wellness management solutions. HealthVault has three major advantages [32]: (a) it presents a low-cost solution in developing and maintaining, (b) it is designed using advanced technology to achieve sustainability, and (c) it offers an easy customization, facilitating programmers to develop customer applications on top of it.

A great number of web and mobile applications that utilize HealthVault are now available. Moreover, many corporations sell portable medical devices (blood pressure monitors, blood glucose meters, pedometers, and more) that take patients' recordings and send them to their HealthVault records. Applications and devices offer a growing range of ways to get the important health information patients that are tracking into their HealthVault record. Many applications let customers analyze and manage that information to help them achieve their health and fitness goals [33].

HealthVault introduces five application connection models in order to help developers decide how their application will integrate with HealthVault [34]: (a) Native HealthVault (Online apps), (b) Linking (Offline apps), (c) Patient Connect, (d) Drop-off/Pick-up, and (e) Software on Device Authentication (SODA).

All these capabilities render HealthVault to be a valuable tool of import and management of health-related information. A major limitation should be stressed; using HealthVault is available only to residents of the United States of America, due to legal obstacles. It is the company's intention to expand its use in other countries, provided that the relevant legal restrictions will be eliminated.

Unlike using it, development of software which supports or uses HealthVault is possible outside the United States as well. For this purpose Microsoft HealthVault Pre-Production Environment (PPE) is available. The PPE is a web server platform that simulates the HealthVault website except that it does not provide access to real user data. It has been designed specifically to support the development of related applications, so that developers can test and evaluate their software. CardioSmart365 uses such an account, created in PPE.

5. CardioSmart365: System Description and Services

5.1. System Architecture. In this work, an integrated system (CardioSmart365) for monitoring and early notification for patients suffering from heart diseases is presented. The system design and implementation use the well-known-service oriented architecture (SOA) to maximize interoperability and scalability, as well as user interface design techniques for optimal presentation. The system consists of web applications, native mobile applications for Smartphones loosely

coupled web services and decision support systems, in order to offer its services to the end users: patients, cardiologists, and general doctors. The four main services the system offers are the following.

(1) Creation and maintenance of the complete electronic medical record (EMR) of patients with heart diseases. EMRs are managed by cardiologists responsible for each patient.

(2) Recording and management of patient's measurements of vital signs performed at home on regular basis, such as blood pressure, blood glucose, oxygen saturation, weight, and height.

(3) Detection of out-of-range measurement values using fuzzy logic and alert firing through a DSS.

(4) Formulation of cardiologic patient profiles based on DSS for the most common heart diseases.

(5) Integrated message management module, for optimal communication between end users and instant notifications. The module also includes automated messages.

The system implements a client-server architecture. Authorised end users have access to the integrated system through client applications, a web application, and a native mobile application for smartphones, with friendly- and easy-to-use interfaces. Great emphasis has been given in the design of user friendly and functional interfaces for both physicians and patients. In particular, the interface of mobile devices is designed in such a way to require the minimum volume of typing data. In order to achieve platform independence, the client applications communicate and exchange data with the database through web services, which allow data interchange through heterogeneous systems. The web services provide functionality with which specific information can be accessed by client applications after authenticated access. CardioSmart365 utilizes the Microsoft HealthVault platform as a backend platform, to store and manage important information of patients' EMRs and measurements, into a uniform format. The system's architecture is shown in Figure 1. Since CardioSmart365 has reached version 1.0 offering the aforementioned services, it will be further developed to incorporate an SMS server component, in order to offer an extra channel of communication between end users, as well as an alternative data transfer method between client and server.

5.2. System Functionality. In this section, we proceed to a thorough analysis of CardioSmart365 functionality and the services it offers to end users (Figure 2). The description of the system's functionality and offered services will be accomplished through the presentation of the following generalized services: (a) vital signs measurement, (b) electronic medical records, (c) cardiologic patient modules, and (d) message management module.

5.2.1. Measurements of Vital Signs. Depending on the severity of patients' health condition, their cardiologists advise them to perform measurements of their vital signs on regular basis. The measurement types may include blood pressure, blood glucose, oxygen saturation, weight, and height, and the frequency of measurements depends, in general, on the patient's health condition. CardioSmart365 offers patients tools to record their measurements through web interfaces and mobile applications (Figure 3) and store them to the system's database. Measurements are performed at home and are imported manually by patients themselves or directly from the medical device (when the device supports connection via Bluetooth or Wi-Fi). Patients can also view measurements of a certain type performed in a specified period of time and create charts. The cardiologist, who is assigned to and responsible for each patient, has access to these measurements for a more comprehensive patient monitoring and decision making. Measurements of vital signs are also stored in Microsoft HealthVault.

An extra control module based on fuzzy sets is developed to check out-of-range measurement values and alert the attending MD. Every attending MD, cardiologist or general practitioner is able to define the fuzzy sets of blood pressure, blood glucose, cholesterol, and weight according to every patient's special needs, achieving in this way a first level of personalization. The fuzzification of blood pressure to *very low, low, normal, little high, medium high,* and *very high* is common to all patients, but the triangular fuzzy sets contain different Universe of Discourses. For example, the universe of discourse for the fuzzy set *little high* can be set from the MD for a patient with a history of a recent heart attack to $(100, 120, 140)$. The same MD can set the same values $(100, 120, 140)$ for the fuzzy set *normal*, in the case of another patient with a different medical history than the previous one, such as a patient with hypertension. A fuzzification and control module is also developed for blood glucose, cholesterol, and weight.

5.2.2. Electronic Medical Record. Through CardioSmart365, cardiologists have access to their patients' medical records. In CardioSmart365, an EMR consists of patient's medical history, medication, laboratory examinations, cardiovascular examinations (physical examinations), periodical measurements of vital signals, and demographics.

The medical history for cardiology (Figure 4) consists of detailed information about coronary artery disease (CAD), intervention for CAD, hypertension, heart failure, valvular heart disease, and heart rhythm disorders. It also includes basic information about cholesterol, diabetes mellitus, tobacco use, family history for heart disease, stroke, peripheral arterial disease, thyroid disease, cancer, lung/liver/kidney/neurological/gastrointestinal/autoimmune/hematologic/endocrine/ophthalmologic/psychiatric disease.

Medication (Figure 5) includes information about patient's current and past medicines, dosage, route, duration, and instructions. Laboratory examinations include information for complete blood count, coagulation times, and biochemical examinations. The cardiovascular examination is a physical examination performed at home or in a clinic, by a general doctor or a cardiologist, and consists of information

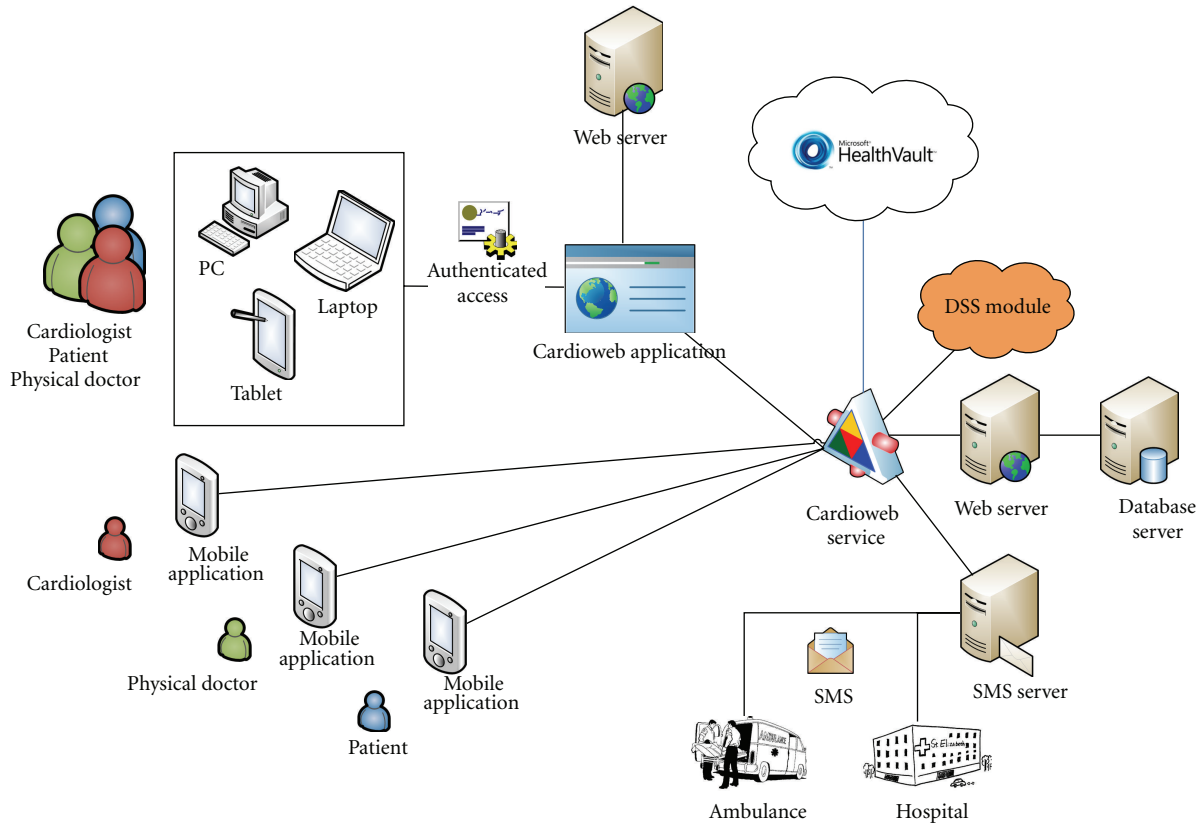


Figure 1: The architecture of CardioSmart365: components interconnection.

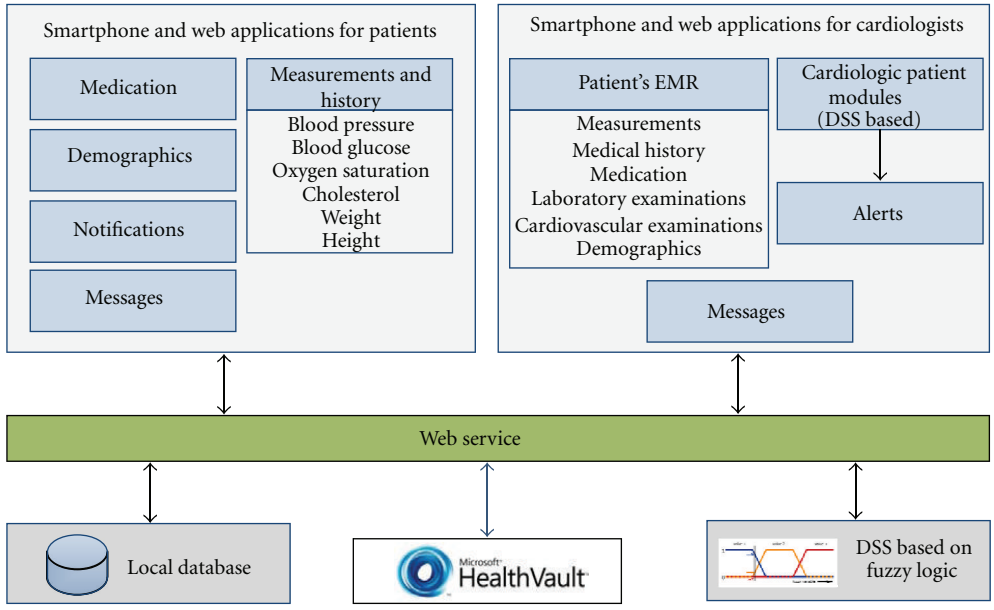


Figure 2: The architecture of CardioSmart365: offered services.

about heart pulse (also femoral and foot pulse), jugular venous pulse, cardiac palpation, and auscultation. The system offers an interface for recording information about the examination the same time it is performed (Figure 6). Measurements of vital signs are performed by patients on regular basis and include blood pressure, blood glucose, oxygen saturation, weight, and height measurements. Demographics include general information for a patient such as name, surname, gender, age, insurance details, contact info and emergency contact details.

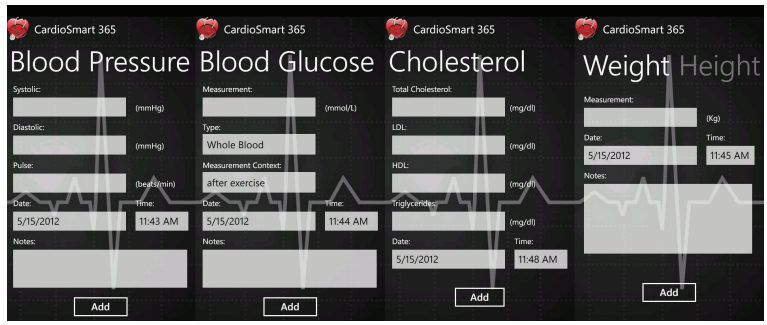


Figure 3: CardioSmart365 mobile interfaces for recording vital signs measurements.

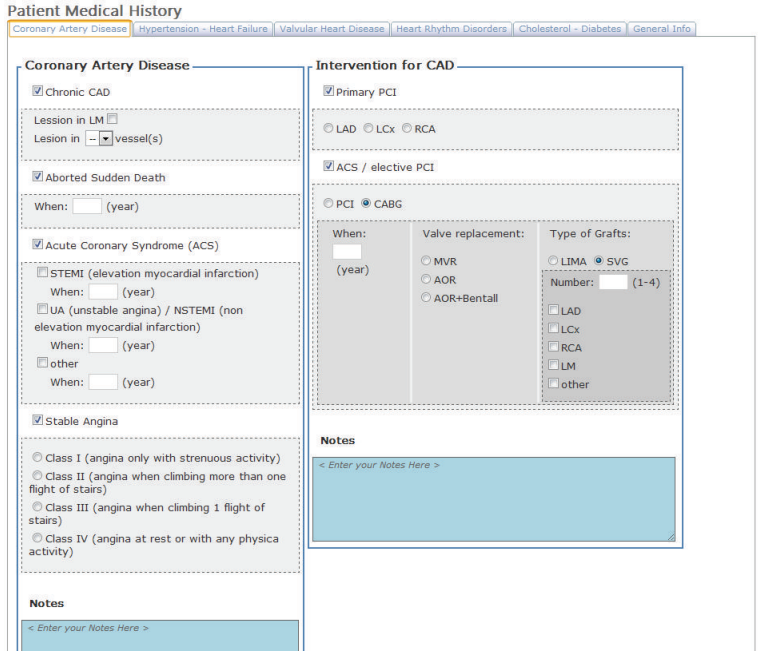


Figure 4: Recording patient's medical history.

A patient's medical history is created in detail the first time the patient visits the cardiologist in his/her office or clinic. Cardiologists have full access to this information and can notify it to general doctors who perform physical examinations in cases of emergency, if it is necessary (read only access permissions). Medication is updated by patients or cardiologists. This information would also be updated directly by pharmacists, as future potential end users of CardioSmart365. Information about laboratory examinations is recorded and stored into CardioSmart365 by patients or cardiologists and in the future directly by clinicians.

Information about a new cardiovascular examination is stored by the cardiologist or general doctor who performs it. Periodical measurements of vital signs are performed at home, are imported by patients, and are available to cardiologists for a more comprehensive patient monitoring and decision making. Demographic information is stored along with the medical history during the patient's first visit to a cardiologist, or by a patient at any time and is updated by patient or cardiologist.

Most of the information described above is also stored in Microsoft HealthVault, including medication, laboratory examinations, periodical measurements, and demographics. This way, the information will be available to third parties after patient's approval. Information characterized as more specific to cardiology, such as the detailed medical history, is stored only in CardioSmart365 database.

5.2.3. Cardiologic Patient Modules. Knowledge from experts, in this case cardiologists, is incorporated in separate modules, focused on the most common heart diseases. The cardiologic patient modules (CPMs) are in fact enhanced cardiologic patient profiles that comprise decision support systems (DSSs) based on fuzzy logic. Fuzzy Logic is used because it offers solutions when a system is so complicated that cannot be mathematically modeled, or when it presents fuzziness. In the case of a cardiology patient and a cardiologist that needs to take therapeutic decisions for him/her, decisions should not be based on rules like the following.

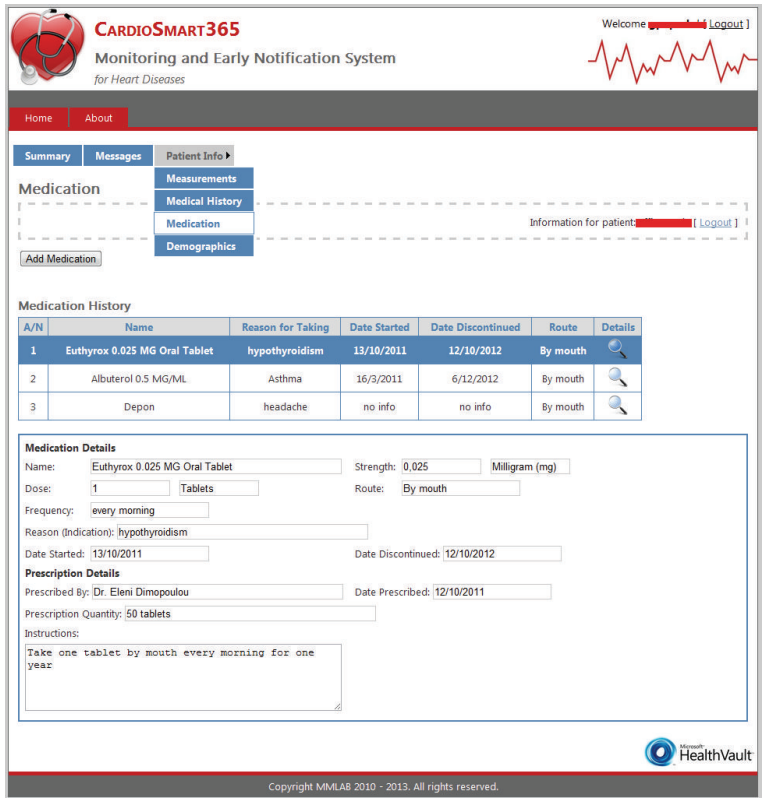


Figure 5: Patient's Medication History and Details from cardiologist's interface.

(1) If *Aortic Stenosis* is 2.5–2.9 m/s AND *Blood Pressure* is 140/90–160/95 mmHg Then *Decision* is . . .

(2) If *Aortic Stenosis* is 3.0–4.0 m/s AND *Blood Pressure* is 140/90–160/95 mmHg Then *Decision* is . . .

It is obvious that different *decisions* result from minor changes. For example, a difference of 1 m/s in the *aortic stenosis* of 2.9 m/s to 3.0 m/s may be the reason for a surgery intervention or not. Therefore, fuzzy sets are designed and fuzzy rules are developed, which imitate better the way MDs think and act. For example, the above decision rules are transformed to fuzzy rules in the form of the following.

(a) If *Aortic Stenosis* is *MILD* AND *Blood Pressure* is *LITTLE HIGH* Then *Decision* is . . .

(b) If *Aortic Stenosis* is *MODERATE* AND *Blood Pressure* is *LITTLE HIGH* Then *Decision* is . . .

The system includes five discrete patient modules:

(I) the coronary artery disease patient module,

(II) the hypertension patient module,

(III) the heart failure patient module,

(IV) the valvular heart disease patient module,

(V) the heart rhythm disorders patient module.

CPMs contain Alert messages and Critical messages automatically arisen when the MD imports data in the DSS. Some decisions from the DSS result deterministically while others require the use of qualitative variables. The qualitative variables are transformed in fuzzy sets used to create fuzzy decision rules. For example, in the valvular heart disease CPM the form of the fuzzy rules is the following.

(1) If *mitral regurgitation* is *ASYMPTOMATIC* AND *LVESd* is *HIGH* then *surgery intervention* is HIGH.

(2) If *aortic regurgitation* is *ASYMPTOMATIC* AND *ef* is *SMALL* then *surgery intervention* is HIGH.

In the Hypertension CPM, Fuzzy Logic is used to help in drugs regulation. The form of the fuzzy rules in this case is:

(1) If *age* is *HIGH* AND *blood pressure* is *LITTLE HIGH* AND *coexisting disease* is Chronic Kidney Disease THEN *Medication* is DIURETIC HIGH.

(2) If *age* is *MEDIUM* AND *blood pressure* is *HIGH* AND *coexisting disease* is Myocardial Infarction THEN *Medication* is B-BLOCKERS HIGH.

Table 1 presents the basic criteria used in the decision mechanism for each one of the five CPMs. Some of the criteria are fuzzified into fuzzy variables, while others are better exploited as crisp variables. The decision includes starting, ending, or configuring medication, limitation of activity, and surgery intervention.

The CPMs are designed to be used from medical doctors, cardiologists and general practitioners. Although

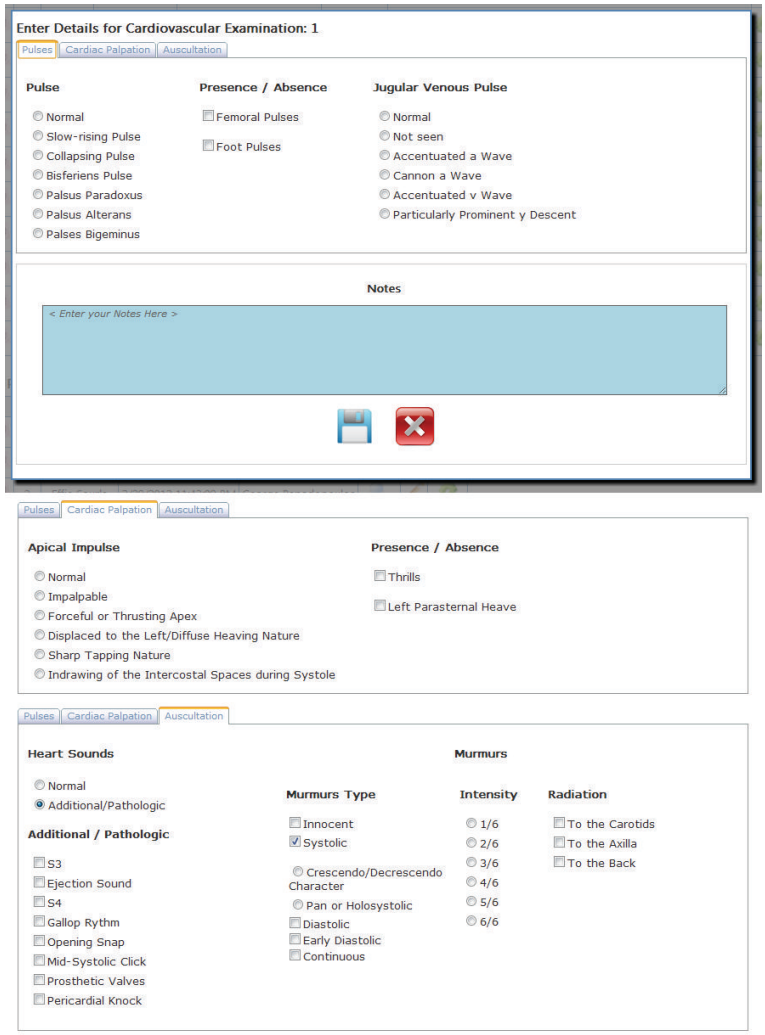


Figure 6: Performing a Cardiovascular Examination.

cardiologists are strongly benefited from the CPMs, general practitioners may find CPMs extremely helpful when they are called to manage a cardiologic patient.

5.2.4. Message Management Module. The message management module is a module of CardioSmart365 that optimizes the communication between end users. The module manages inbox and sent messages of all end users and sends automated messages when necessary. The automated messages are formed by customizing specific message templates stored in the system's database. When an automated message is sent or when one user sends a message to another, three actions take place; a message is created internally in the system database and is visible to all recipients, an email is sent to the recipients who have registered their email addresses, and an SMS is sent to the recipients who have registered the numbers of their mobile phones. Through the messaging module interface (Figure 7), the end user can view his/her inbox, edit and send outbox messages, create, save, or send a new message and delete messages.

5.3. Software Framework. The multimodule software architecture is developed in Visual Studio 2010, using the programming language C#. The web services implementation follows the Windows Communication Foundation (WCF) framework that provides a unified programming model for rapidly building service-oriented applications. The mobile application is implemented in Windows Phone 7, the new Microsoft platform for smartphones, specifically version 7.5 Mango, which is an upgraded version of the initial one. The Silverlight toolkit for Windows Phone was used, which allows rapid creation of Rich Internet Application-style user interfaces. The integrated system's database is developed in SQL Server 2008.

A prerequisite for developing applications interconnecting to Microsoft HealthVault is the use of the corresponding library [25]. The library enables encrypted communication among the application and the server of HealthVault. The respective SDK provides supplementary tools for interaction with the platform, instructions for the programmer as well as some standard applications. Each application is represented by a unique ID, which is received through HealthVault

Table 1: Cardiologic patient modules and decision support.

Cardiologic patient module	Fuzzy variables	Crisp variables	Decision
Coronary artery disease	Class Duration	Category	Medication
Hypertension	Age Blood pressure	Coexisting disease	Medication
Heart failure	Class	Cause	Limitation of activity
Valvular heart disease	Mitral regurgitation Aortic regurgitation LVESd ef		Surgery intervention
Heart rhythm disorders	Rate control Rhythm control	Type	Medication

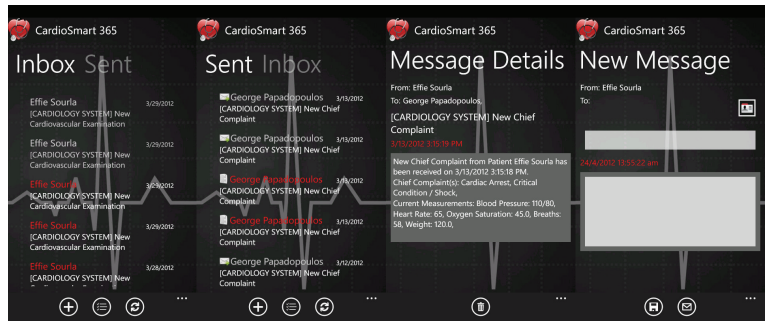


Figure 7: The mobile interface of the Message Management Module.

Application Configuration Center. The DSSs based on fuzzy logic are developed in Matlab® Fuzzy Logic Toolbox.

6. Added Value

CardioSmart365 impact is separated to benefits for five categories, which are thoroughly described below:

Every day clinical practice concerns MDs, nurses, hospital staff, outhospital healthcare organisations, and patients. CardioSmart365's patient modules, forms, and DSS can be used as guidelines that are followed for patients with a specific cardiologic disease, standardising in this way the tasks that have to be done for every patient, each day and by whom. In this way, better collaboration is established between all the involved working teams inside and outside the hospital. In addition, the direct interconnection and fast briefing, between different health sectors supported from the system, reduces fraud and optimizes emergency incidents process management.

Medical doctors many times need to act under stress and to take decisions in a fast way. Therefore, speed is important for MDs. cardioSmart365 is a "doctor-friendly" tool that helps recording data in a systematic and quick way, therefore, helping MDs to work faster and in more safety. In addition, CardioSmart365 benefits young cardiologists and general practitioners (GPs) as an educational and decision support tool, since it provides patient specific guidance in the rapprochement of CPs.

Cardiologic patients dispose a customised EHR. The benefits of EHRs are widely known, but the customised EHR for CPs contains detailed data, formatted in a way that can be easily understood and quickly viewed even from third parties MDs after years. After all, patients suffering from cardiologic diseases, most of the time in reflect chronic pathologies that require medication and followup for the rest of their life. This way, patients receive lifelong advantageous healthcare.

Research and Science strongly benefited towards better monitoring and understanding of cardiologic diseases. CardioSmart365 is a tool for recording and studying scientifically validated data elements of cardiologic diseases. Moreover, CardioSmart365, in the administrator level, can correlate and conclude upon data recorded not only from different places of the same region or country, but from different countries and continents, because it can be accessed from any personal computer connected to the Internet. In this way, maybe for the first time so easily, detailed data from all over the world can be gathered in one database, continuously updated for scientific research.

Healthcare systems are interested in estimating health related economical costs. CardioSmart365 can be used for a reliable estimation of the economical cost that a patient encumbered a healthcare system. CardioSmart365 stores

data concerning the examinations of a patient that their cost is usually the higher cost that burdens a healthcare system. Additionally, CardioSmart365 can be used for providing estimations on the cost of every cardiologic disease per patient and to estimate annual costs for every new diagnosed patient.

7. Future Steps

CardioSmart365 was developed to offer advanced added value services for cardiologic patients and physicians into a fast, flexible, and easy-to-use integrated solution. However, many steps towards optimizing the system have to be done. Some direct future evolvements concern (a) healthcare systems and (b) research and science.

7.1. Healthcare Systems. CardioSmart365 is designed to offer an easy way for the incorporation of economic costs customized to different healthcare systems. This future module, apart from the already provided costs of laboratory examinations, will provide costs related to medication, MDs rewards, patients' transportations, and other economic issues important for healthcare systems around the world. Moreover, tools for estimation and various indices used for better monitoring and prediction of health involved costs will be established. For example, questions like "What is the annual cost for a Healthcare system for a patient with a coronary artery disease?" will be answered, or predictions like "How many new diagnosed patients with a coronary artery disease are expected for the upcoming year" will be estimated. AI methods will be closely studied due to the help that they provide in prediction problems.

7.2. Research and Science. Using the web applications and services of CardioSmart365, feedback from institutional centers specialized on cardiologic diseases will be collected and incorporated to future versions of CardioSmart365. The cardiologic patient modules will be continuously updated in an automated way through the tools that will be developed. Knowledge from experts will be further continuously incorporated to the DSS of CardioSmart365, optimizing their support to MDs, leading towards personalized patient profiles and personalized medicine. In this way specialized knowledge and protocols from pioneers in the field of cardiology will be spread all around the world in a direct, fast, and easy-to-follow way.

CardioSmart365 will further adopt clinical data and data involved in healthcare in a greater detail. As a first attempt to collect and group the basic data elements sufficient for a cardiology patient, MDs and the rest staff involved in Healthcare systems, CardioSmart365 will incorporate serious feedback from its end users for optimization and further evolvements.

8. Conclusions

An integrated system based on web applications, Smartphones, and an interconnection to Microsoft HealthVault platform is developed, for (a) monitoring chronic cardiology patients and (b) early notifying and optimizing the process management of an emergency cardiologic incident. The system supports cardiologic patient modules based on common cardiology diseases and customized DSS based on fuzzy logic. The benefits of the system concern patients, MDs, everyday clinical practice, research and science, and healthcare systems.

Acknowledgments

This research has been cofinanced by the European Union (European Social Fund—(ESF)) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)—Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

[1] I. Varlamis and I. Apostolakis, "Medical informatics in the Web 2.0 era," in *Proceedings of the 1st International Symposium on Intelligent Interactive Multimedia Systems and Services*, Piraeus, Greece, July 2008.

[2] E. Faliagka, V. N. Syrimpeis, A. Tsakalidis, G. K. Matsopoulos, J. Tsaknakis, and G. Tzimas, "Diagnosis: A global alignment and fusion medical system," in *Proceedings of the 3rd International Conference on Health Informatics, HEALTHINF 2010*, pp. 21–28, esp, January 2010.

[3] C. H. Liu, Y. F. Chung, T. W. Chiang, T. S. Chen, and S. D. Wang, "Mobile agent approach for secure integrated medical information systems," *Journal of Medical Systems*, vol. 36, no. 5, pp. 2731–2741, 2012.

[4] I. Babaoglu, O. K. Baykan, N. Aygul et al., "A comparison of artificial intelligence methods on determining coronary artery disease," in *Advances in Information Technology*, Communications in Computer and Information Science, pp. 18–26, Springer, Berlin, Germany, 2010.

[5] A. M. Lindquist, P. E. Johansson, G. I. Petersson, B. I. Saveman, and G. C. Nilsson, "The use of the Personal Digital Assistant (PDA) among personnel and students in health care: a review," *Journal of Medical Internet Research*, vol. 10, no. 4, article no. e31, 2008.

[6] M. Prgomet, A. Georgiou, and J. I. Westbrook, "The impact of mobile handheld technology on hospital physicians' work practices and patient care: a systematic review," *Journal of the American Medical Informatics Association*, vol. 16, no. 6, pp. 792–801, 2009.

[7] V. Gkintzou, T. Papablasopoulou, V. Syrimpeis, E. Sourla, G. Tzimas, and A. Tsakalidis, "A web and smart phone system for tibia open fractures," in *ENTERprise Information Systems*, J. Varajão, P. Powell, and R. Martinho, Eds., vol. 221 of *Communications in Computer and Information Science*, pp. 413–422, Springer, Berlin, Germany, 2011.

[8] V. L. Patel, E. H. Shortliffe, M. Stefanelli et al., "The coming of age of artificial intelligence in medicine," *Artificial Intelligence in Medicine*, vol. 46, no. 1, pp. 5–17, 2009.

[9] Y. Peng, Y. Zhang, and L. Wang, "Artificial intelligence in biomedical engineering and informatics: an introduction and review," *Artificial Intelligence in Medicine*, vol. 48, no. 2-3, pp. 71–73, 2010.

[10] T. P. Exarchos, M. G. Tsipouras, C. P. Exarchos, C. Papaloukas, D. I. Fotiadis, and L. K. Michalis, "A methodology for the

automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree," *Artificial Intelligence in Medicine*, vol. 40, no. 3, pp. 187–200, 2007.

[11] S. N. Ghazavi and T. W. Liao, "Medical data mining by fuzzy modeling with selected features," *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 195–206, 2008.

[12] M. A. Grando, M. Peleg, M. Cuggia, and D. Glasspool, "Patterns for collaborative work in health care teams," *Artificial Intelligence in Medicine*, vol. 53, no. 3, pp. 139–160, 2011.

[13] M. Rowan, T. Ryan, F. Hegarty, and N. O'Hare, "The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors," *Artificial Intelligence in Medicine*, vol. 40, no. 3, pp. 211–221, 2007.

[14] A. Chu, H. Ahn, B. Halwan et al., "A decision support system to facilitate management of patients with acute gastrointestinal bleeding," *Artificial Intelligence in Medicine*, vol. 42, no. 3, pp. 247–259, 2008.

[15] X. Zhou, S. Chen, B. Liu et al., "Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support," *Artificial Intelligence in Medicine*, vol. 48, no. 2-3, pp. 139–152, 2010.

[16] S. Chemlal, S. Colberg, M. Satin-Smith et al., "Blood glucose individualized prediction for type 2 diabetes using iPhone application," in *Proceedings of the 37th Annual Northeast Bioengineering Conference, (NEBEC'11)*, April 2011.

[17] B. Silva, I. Lopes, J. Rodrigues, and P. Ray, "SapoFitness: a mobile health application for dietary evaluation," in *Proceedings of the IEEE 13th International Conference on e-Health Networking, Applications and Services*, pp. 375–380, June 2011.

[18] F. Sposaro and G. Tyson, "iFall: An android application for fall monitoring and response," in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, (EMBC'09)*, pp. 6119–6122, September 2009.

[19] F. Sposaro, J. Danielson, and G. Tyson, "IWander: an Android application for dementia patients," in *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (EMBC'10)*, pp. 3875–3878, September 2010.

[20] M. N. K. Boulos, S. Wheeler, C. Tavares, and R. Jones, "How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX," *BioMedical Engineering Online*, vol. 10, article 24, 2011.

[21] N. Houston, "The best medical Iphone apps for doctors and med students," October 2010, http://blog.softwareadvice.com/articles/medical/the-best-medical-iphone-apps-for-doctors-and-med-students-1100709/.

[22] QxMD, 2012, http://www.qxmd.com/specialty/medicine/cardiology-medical-apps-iphone-blackberry-android.

[23] Webicina—Cardiology in Social Media, 2012, http://www.webicina.com/cardiology/cardiology-mobile-applications/.

[24] Hearth360 Cardiovascular Wellness Center, May 2012, https://www.heart360.org/Default.aspx.

[25] Microsoft HealthVault Development Center, May 2012, http://msdn.microsoft.com/en-us/healthvault/default.aspx.

[26] G. Guidi, E. Iadanza, M. Pettenati et al., "Heart failure artificial intelligence-based computer aided diagnosis telecare system," in *Impact Analysis of Solutions for Chronic Disease Prevention and Management*, M. Donnelly, C. Paggetti, C. Nugent, and M. Mokhtari, Eds., vol. 7251, pp. 278–281, Springer, Berlin, Germany, 2012.

[27] J. W. Wallis, "Invited commentary: use of artificial intelligence in cardiac imaging," *Journal of Nuclear Medicine*, vol. 42, no. 8, pp. 1192–1194, 2001.

[28] P. C. Tang, J. S. Ash, D. W. Bates, J. M. Overhage, and D. Z. Sands, "Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption," *Journal of the American Medical Informatics Association*, vol. 13, no. 2, pp. 121–126, 2006.

[29] NAHIT Report, April 2008, http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_10741_848133_0_0_18/10_2_hit_terms.pdf.

[30] A. Sunyaev, D. Chornyi, C. Mauro, and H. Krcmar, "Evaluation framework for personal health records: Microsoft HealthVault vs. Google Health," in *Proceedings of the 43rd Annual Hawaii International Conference on System Sciences, (HICSS'10)*, usa, January 2010.

[31] A. Sunyaev, A. Kaletsch, and H. Krcmar, "Comparative evaluation of Google Health API vs Microsoft HealthVault API," in *Proceedings of the 3rd International Conference on Health Informatics, (HEALTHINF'10)*, pp. 195–201, Valencia, Spain, January 2010.

[32] L. Liao, M. Chen, J. J. Rodrigues, X. Lai, and S. Vuong, "A novel web-enabled healthcare solution on healthvault system," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1095–1105, 2012.

[33] Microsoft HealthVault, Apps and Devices, May 2012, http://www.microsoft.com/en-us/healthvault/tools-devices/overview.aspx.

[34] S. Nolan, "Microsoft HealthVault Application Connection Recommendations," 2010 http://download.microsoft.com/download/7/4/E/74EA8944-199C-4F56-B3BB-810586942-5BC/HealthVault%20Application%20Integration%20Recommendations%20v1.pdf.

Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2012, Article ID 457351, 7 pages
doi:10.1155/2012/457351

Research Article

A Cultural Algorithm for the Representation of Mitochondrial Population

Athanasios Alexiou and Panayiotis Vlamos

Department of Informatics, Ionian University, Plateia Tsirigoti 7, 49100 Corfu, Greece

Correspondence should be addressed to Athanasios Alexiou, alexiou@ionio.gr

Received 31 May 2012; Accepted 14 August 2012

Academic Editor: Catalina Cocianu

We propose a novel Cultural Algorithm for the representation of mitochondrial population in mammalian cells as an autonomous culture. While mitochondrial dysfunctions are highly associated with neurodegenerative diseases and related disorders, an alternative theoretical framework is described for the representation of mitochondrial dynamics. A new perspective of bioinspired algorithm is produced, combining the particle-based Brownian dynamics simulation and the combinatorial representation of mitochondrial population in the lattice, involving the optimization problem of ATP production in mammalian cells.

1. Introduction

Considering the latest researches, disruptions in the regulation of mitochondrial dynamics, low-energy production, increased reactive oxygen species, and mtDNA damage are relevant to human diseases, mainly in neurogenerative diseases and cancer. Recent discoveries have highlighted that neurons are reliant particularly on the dynamic properties of mitochondria. In addition, mitochondria are actively recruited to subcellular sites, such as the axonal and dendritic processes of neurons. Defects in mitochondrial dynamics are associated with neurodegenerative disease. For example, Charcot-Marie-Tooth type 2A, a peripheral neuropathy, and dominant optic atrophy, an inherited optic neuropathy, result from a primary deficiency of mitochondrial fusion. Moreover, several major neurodegenerative diseases including Parkinson's, Alzheimer's, and Huntington's diseases involve disruption of mitochondrial dynamics.

On the other hand, cultural algorithms are a class of population concepts, principles, mechanisms, and optimization techniques that work on a principle inspired by nature: evolution of species [1]. These algorithms are very useful tools in a large number of applications in optimization, control, signal processing, or machine learning [2].

Lately, researchers attempted to model the cultural evolution process from both a microevolutionary perspective in terms of the transmission of behaviors or traits between individuals in a population and a macroevolutionary perspective in terms of the formation of generalized beliefs based upon individual experiences. These generalized beliefs can serve to constrain the behaviors of individuals within the associated population [3].

According to Reynolds, Cultural Algorithms are a class of computational models of cultural evolution that support such a dual inheritance perspective. This approach provides a framework in which to describe all of the current models of cultural evolution from a computational point of view since any of the single inheritance systems can be produced as a special case [3].

Cultural Algorithms, are based on the supposition that one can get better learning rates for an evolutive genetic algorithm [4] adding to it one more element of evolutive pressure—called Belief Space, a mechanism of cultural pressure. Therefore, a system of double inheritance, both genetic and cultural, could better respond to a large number of problems, while cultural evolution enables societies to evolve or adapt to their environments at rates that exceed that of biological evolution based upon genetic inheritance alone [5, 6].

In this paper we investigate the application of Cultural Algorithms on the representation of mitochondrial population in mammalian cells, a biological system with high

complexity, different capabilities and operations, and great importance for the human health.

2. The Mitochondrial Population and Dynamics

The number of mitochondria in a cell is regulated to match the cell's requirements for ATP, while fusions and fissions play a functional role in maintenance of proper inner membrane electrical potential. Mitochondria provide most of the ATP for cellular reactions. ATP production in mitochondria is coupled to an electron transport system in which the passage of electrons down the various electron carriers is associated with the transport of protons from the matrix into the intermembrane space. The majority of these protons reenter the mitochondrial matrix by the ATP syntheses, thereby generating ATP.

Mitochondria are involved in numerous metabolic and cellular processes [7]. Besides the citric acid cycle and the oxidative phosphorylation, these processes also include the urea cycle and the oxidation of fatty acids. Other reactions carried out and orchestrated are the biosynthesis of heme, several amino acids, and vitamin cofactors, as well as the formation and export of iron-sulphur clusters [8]. Beyond these metabolic functions, mitochondria are also involved in the programmed cell death [9] and in case of their dysfunction in ageing [10] and several diseases [11].

2.1. Mitochondrial Structure. Found in most eukaryotic cells, mitochondria are subcellular organelles that play a central role in energy metabolism. The key feature of the mitochondrion is the presence of two membranes that encapsulate a protein-rich central matrix. Specifically, this organelle is compartmentalized by two membranes into four compartments. A smooth outer membrane surrounds and isolates the organelle from the cytosol while the inner membrane with several invaginations, called cristae, divides them further into the mitochondrial intermembrane space and the matrix. The organization of the inner membrane was dissected in recent studies using improved electron microscopic and tomographic techniques [12, 13]. The inner membrane contains the protein complexes and redox cofactors involved in electron transfer and ATP synthesis.

Besides the outer membrane, an inner boundary membrane is connected by several tubular junctions to the cristae, creating a distinction between the intermembrane and the intercristal space. This basic concept of the inner membrane is structurally dynamic with respect to cristae connection to each other or with the inner membrane and can be considerably varied among different organisms' tissues or physiological conditions [12]. The mitochondrial matrix contains a highly concentrated mixture of enzymes involved in all aspects of metabolism, in addition to the mitochondrial genome which, in mammalian mitochondria, encodes 39 genes involved in mitochondrial function. Families of mitochondrial carriers, of which approximately 50 have been identified in the human genome, are present to enable exchange between the intermembrane space and the matrix [14].

The fact that many central processes in eukaryotic cells are functionally linked to this double membrane shielded organelle requires an interface between the mitochondrial compartment and the cytoplasm. To achieve this intracellular exchange the two membranes include a variety of transport and receptor proteins [15–17] as well as a specific subset of translocases involved in the import and assembly of mitochondrial proteins [18].

2.2. Mitochondrial Processes: Fusion, Fission, Motility, and Mitophagy. The number of mitochondria in a cell is regulated to match the cell's requirements for ATP, while fusions and fissions play a functional role in maintenance of proper inner membrane electrical potential. Without the mitochondrial dynamics, the mitochondrial population consists of autonomous organelles that have impaired function [19]. In a wild type cell, high rates of fusion and fission are independent events, which constantly change the identity of individual mitochondria, as well as the motility and the mitophagy.

An individual mitochondrion is not an autonomous organelle. The hundreds of mitochondria within a typical cell undergo continual cycles of fusion and fission. Because mitochondria have an outer lipid membrane as well as an inner one, each fusion event requires the coordinated fusion of the membranes [20, 21]. Fusion is likely to protect function by providing a chance for mitochondria to mix their contents, thus enabling protein complementation, mtDNA repair, and equal distribution of metabolites, helping the isolation of damaged-mitochondrial segments and promoting their autophagy [22, 23]. In contrast, fission acts in order to facilitate equal segregation of mitochondria into daughter cells during cell division and to enhance distribution of mitochondria along cytoskeletal tracks [24].

Intuitively, it is easy to imagine how mitochondrial fusion and fission can change the morphologic characteristics of mitochondria [20–22]. Fusion results in fewer and longer mitochondria, whereas fission results in more and shorter mitochondria. Indeed, genetic studies indicate that cells with mutations in the genes required for mitochondrial fusion have fragmented mitochondria instead of the tubular mitochondrial network observed in normal cells [24]. Similarly, cells with mutations in genes required for mitochondrial fission have excessively elongated and interconnected mitochondria because of unopposed fusion. Peroxisomal shape is also controlled by fission; the role of fusion is less clear [20].

Additional studies have also shown that mitochondrial fission precedes apoptosis [25]. Defects in mitochondrial fusion cause neurodegenerative disease [23, 24]. Charcot-Marie-Tooth disease type 2A, an autosomal dominant neuropathy of long peripheral nerves, is caused by mutations in MFN2. Moreover, dominant optic atrophy, the most commonly inherited optic neuropathy, is caused by mutations in OPA1. This apparent sensitivity of neurons to defects in mitochondrial dynamics probably depends on the special requirements of neurons for mitochondrial function [20]. An ultrastructural hallmark of the synapse is the presence of abundant mitochondria, which maintains calcium homeostasis and levels of ATP production that,

in turn, are critical to nerve transmission. Neurons have extraordinarily long cellular processes, and tight control of mitochondrial dynamics is probably necessary for distributing active mitochondria to dendrites and axon terminals. Given the important roles of mitochondrial dynamics in human physiologic processes, it would not be surprising to find additional diseases caused by mutations in genes that control mitochondrial fusion and fission [20].

Another aspect of mitochondrial dynamics beyond fusion and fission is the motility of mitochondria [21]. This aspect is critically important in highly polarized cells, such as neurons [26], which require mitochondria at sites distant from the cell body, but can also be crucial to cellular function in smaller cells [27]. Defects in both fusion and fission have been shown to decrease mitochondrial movement. In neurons lacking mitochondrial fusion, both increased mitochondrial diameter due to swelling and aggregations of mitochondria seem to block efficient entry into neurites, resulting in a dearth of mitochondria in axons and dendrites [28]. These defects result in improperly developed neurons or gradual neurodegeneration.

Autophagy is a mechanism whereby eukaryotic cells degrade their own cytoplasm and organelles [29]. Autophagy functions as a homeostatic nonlethal stress response mechanism for recycling proteins to protect cells from low supplies of nutrients and as a cell death mechanism. This degradation of organelles and long-lived proteins is carried out by the lysosomal system; thus, a hallmark of autophagy is accumulation of autophagic vacuoles of lysosomal origin [23]. Autophagy has been seen in developmental and pathological conditions. Mitophagy denotes the degradation of mitochondria through autophagy. Although the existence of mitophagy has been known for some time, it has been unclear whether mitochondria are randomly or selectively targeted for mitophagy [21].

3. An Alternative Biological Culture

3.1. Mitochondrial Civilization Associated with Neurogenerative Diseases. From a geometrical point of view, a number of (un)correlated factors can affect the mitochondrial shape. This illustrates a more complex problem, while morphological changes in mitochondrial structure are associated with biological dysfunctionalities and electrophysiology problems [30]. These effects are directly or indirectly correlated with human neurodegenerative diseases. While fusions and fissions contribute to the wide variety of mitochondrial morphologies, a discrete mitochondrion at one point in time will be changed at a later time by the addition of new mitochondrial material through fusion or by the removal of material through division. It is a logical consequence of high probability that after a certain period of successful events (fusions and fissions) the inner structure will totally lose its initial characteristics in a nonreversible way, restricting the inner space and reducing the corresponding area and energy [30]. It is obvious that any failure in inner membrane mitochondrial fissions can easily generate unstable electric potential, effecting functionality and reduce voltage gradient. Fusion and fission seems to be required to maintain mitochondrial function, as independent and different mechanisms. Fusion is likely to protect function by providing a chance for mitochondria to mix their contents, thus enabling protein complementation, mtDNA repair, and equal distribution of metabolites, helping the isolation of damaged mitochondrial segments and promoting their autophagy. In contrast, fission acts in order to facilitate equal segregation of mitochondria into daughter cells during cell division and to enhance distribution of mitochondria along cytoskeletal tracks. The failure in this biological machinery may also promote apoptosis.

Even though, these four processes are independent, it is clear that any interactions will be critically important in neurons. For example, defects in both fusion and fission have been shown to decrease mitochondrial movement. The large tangle of highly interconnected mitochondria in fission-deficient cells prevents efficient movement, especially into small pathways such as neuronal processes. While mitophagy denotes the degradation of mitochondria through autophagy, recent findings indicate that mitophagy can selectively degrade defective mitochondria.

Especially in the case of Alzheimer's Disease, scientists used brain tissue from cases with a diagnosis of AD [31, 32], as well as control cases with no clinical or pathological history of neurological disease, applying cytological *in situ* hybridization, immunocytochemistry, and morphometry [33] techniques, showing that the area of intact mitochondria is significantly decreased in AD. While AD can be genetically classified as familiar or sporadic, researchers proposed that the case of sporadic AD is not caused by the accumulation of amyloid-β (Aβ), but instead is a consequence of a decline in mitochondrial function with age [34, 35]. Additionally, the overexpression of Aβ causes an alteration in the mitochondrial fission and fusion proteins resulting in mitochondrial dysfunction, mitochondrial fragmentation, increase in reactive oxygen species (ROS) and ATP production, and reduced mitochondrial membrane potential [36].

3.2. A New Cultural Algorithm. From the philosopher Aristotle to the anthropologist Geertz [37], any civilization consists of citizens, where besides the variation of their genetic characteristics, they participate to the social evolution through their behaviour and several others mechanisms, rules, principles, or even more principles.

Additionally researchers [38] introduce the term Artificial Culture where, a new knowledge domain tries to connect the models found in complex adaptive systems to the models found in the domain of culture. According to this terminology Artificial Culture is a population of individual agents, with its own sense, with its own cognition and performance, interacting in a social ambient with others agents in a physical environment of artifacts and others objects [38].

According to Holland [39], a complex system usually has the following characteristics, which obviously in the case of mitochondrial populations are adapted in most of the cases.

(i) Relationships in complex system are nonlinear.

(ii) Complex systems contain feedback loops.

$t \leftarrow 1$
Generate Civilization: N individuals are distributed
in the parametric space, assuming that correspond
to a local martingale
Initialize Population P(t)
Initialize Belief Space BS(t)
While (ATP Production = Acceptant) and (t < constant value) do
{
Evaluate P(t)
Evolve ((P(t), *Merging*(P(t), *Combine*(P(t)))), *Influence*(BLF(t)))
Vote (BS(t), *Accept*(P(t)))
Evaluate (P(t), *Adjust*(BS(t)))
Update (BS(t), *Accept*(P(t)))
$t \leftarrow t + 1$
Select P(t) from P($t - 1$)
}
End

Algorithm 1: Cultural Algorithm.

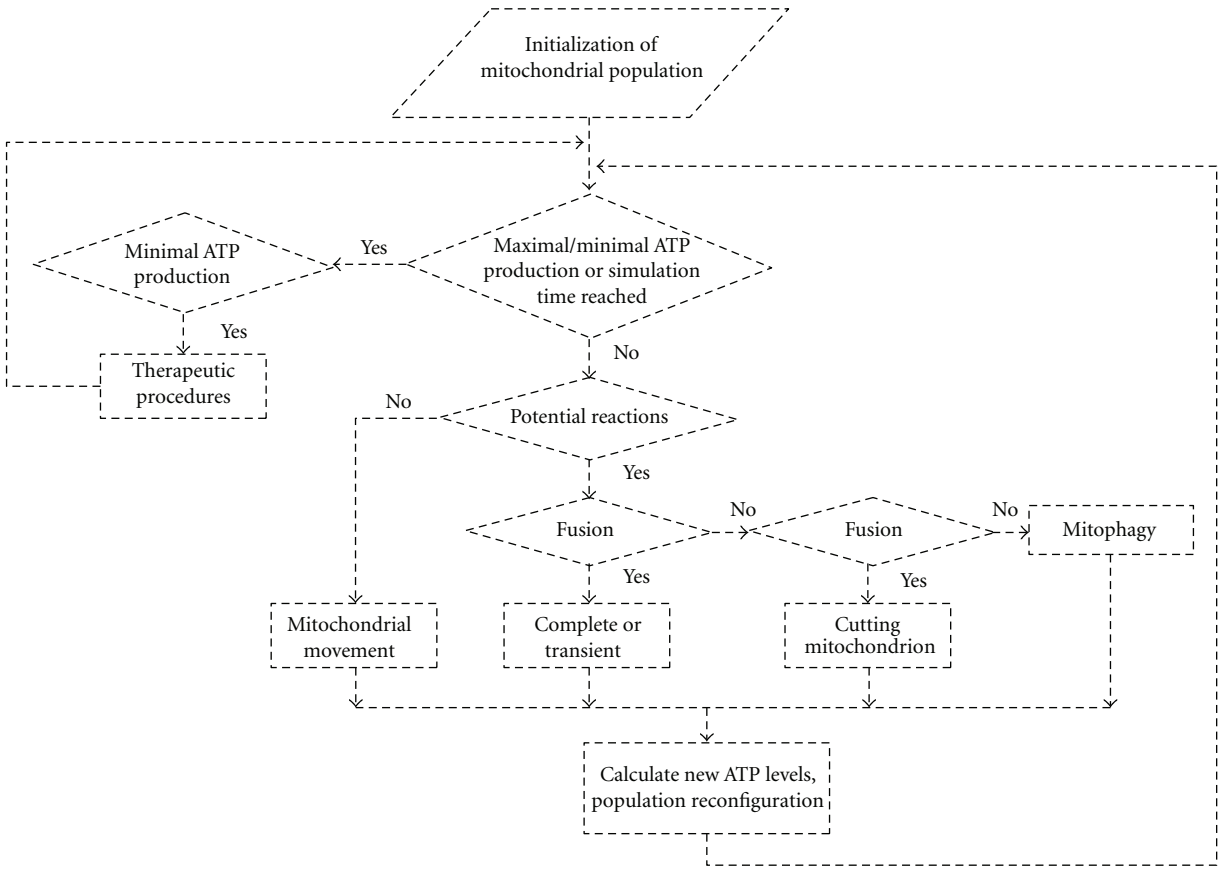


Figure 1: The Cultural Algorithm for the representation of mitochondrial population.

(iii) Complex systems are open.

(iv) Complex systems have memory (even though mitochondria seem to obey the memoryless phenomenon [30]).

(v) Complex system may produce emergent phenomena.

The proposed Cultural Algorithm for the representation of the mitochondrial population is an extended version of Reynolds [3], adapted in mitochondrial terminology (operations) where P(t) represents the population and BS(t) the Belief Space at time t.

The algorithm starts with the generation of the civilization, where the initialization of both the Population and

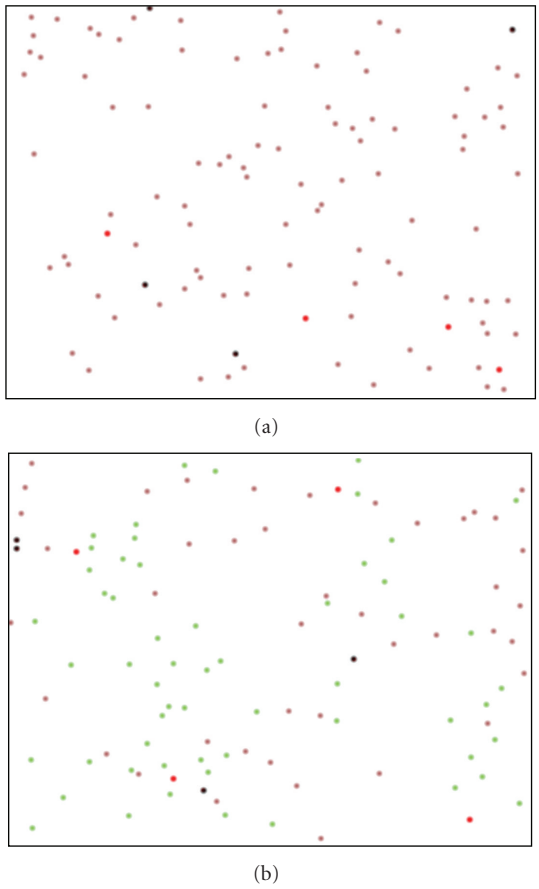


Figure 2: (a) Initialization of population. (b) Simulation results after N Steps.

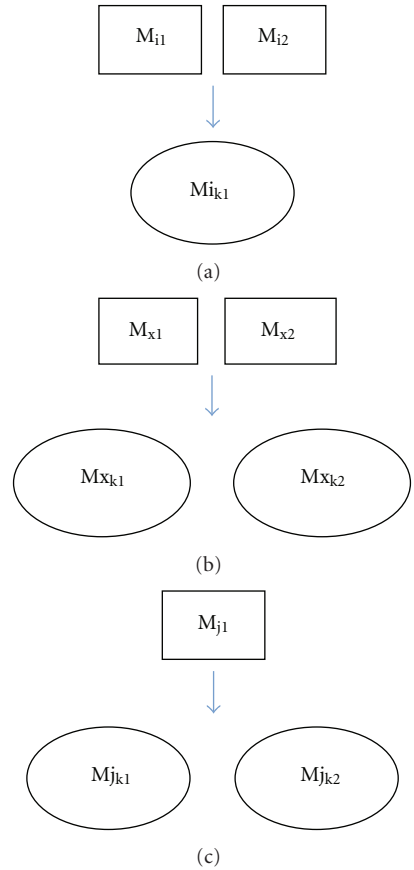


Figure 3: (a) Complete fusion. (b) Transient fusion. (c) Fission.

the Belief Space, occurs. Due to the biological identities of mitochondria in mammalian cells, and their stochastic distribution in the lattice space, we assumed that they produce a local martingale limiting their ATP production within the normal measurement limits. The proposed algorithm enters the evolution loop until the ATP production is decreased in a level likely to result in a neurological disorder. At the beginning of each generation, individuals in the Population Space are first evaluated and then interact to each other through the operation factors "Merging" and "Combine" (representing in a way the mitochondrial dynamics and the communication paths between them in order to exchange their contents) resulting a new selection of individuals through the Influence function.

Additionally, the two feedback functions Accept and Influence, give the opportunity to the population component and the Belief Space to interact with each other in the way that human culture evolve [40–42].

The pseudocode of the proposed bioinspired algorithm is given in Algorithm 1.

By the mechanism of mitochondrial dynamics these organelles can undergo exchange their contents for desired morphology and topology. The above algorithm is a more sophisticated control of mitochondrial proper functioning, setting as population belief a total optimum in the cell space.

Empirically, fusion-deficient mitochondria display loss of directed movement, instead hovering in a manner reminiscent of Brownian motion [43]. It is well known that particle-based Brownian dynamics simulations offer the opportunity to not only simulate diffusion of particles but also the reactions between them. Particle-based simulations naturally incorporate the concepts of space, crowding, and stochasticity [44, 45]. Those methods treat proteins or other reactants explicitly, and the time-evolution of particle positions is sampled at discrete time intervals by Brownian dynamics simulations [46, 47]. The basic motion of a particle undergoing diffusion can be described by Einstein's diffusion equation [48] as follows:

$$\frac{\partial}{\partial t}P(r,t \mid r_0,t_0) = D\nabla^2 P(r,t \mid r_0,t_0), \tag{1}$$

where $P(r,t \mid r_0,t_0)$ is the probability that the particle will be at position r at time t given the particle was initially at position r_0 at time t_0. The rate of diffusion is given by the parameter D. When the diffusive motions of multiple particles are simulated, in a traditional Brownian dynamics algorithm the distribution of particle displacements $P(r,t)$ is sampled for each particle every time step [49].

Algorithm 1 is a a novel theoretical approach in order to represent mitochondrial population of mammalian cells as an independent culture. We assigned a random walk in

these subcellular compartments of a neural cell as follows: viruses to destroyed mitochondria, infected individuals to mitochondria with decreased ATP production, and uninfected individuals to healthy mitochondria.

A more informative representation of the above pseudocode can be seen in Figure 1, where the main feature involves the optimization level of ATP production. In this Cultural Algorithm the stochastic evolution function is performed until the solution has reached a total optimum, the maximal or minimal ATP level, responsible for several human diseases.

We tested our algorithmic representations in a multiagent simulation environment called MASON, which is a single-process discrete-event simulation core and visualization library for biological systems [50].

The population is represented by the colour dots [50], brown for the healthy mitochondria, green for the mitochondria with decreased functionality, and the red for the destroyed mitochondria. There are also some cases of healthy mitochondria, which are represented with black dots with red cross, where destroyed mitochondria can be disinfected through the operation of fusion (Figure 2).

It is obvious that the formulation of mitochondrial population as a biological culture, gives us the potentiality of visualization of the two main observed classes of fusion events in mammalian cells. The complete fusion and the transient fusion events [51], wherein two mitochondria came into close apposition, exchanged soluble intermembrane space and matrix proteins, and reseparated, preserving the original morphology.

In this phase, we are more focused in the three main mitochondrial operations: the complete fusion, the transient fusion of two mitochondria, and the fission of one mitochondrion (Figure 3).

It is obvious that further simulation testing of these dependent operations will give us the opportunity of modelling efficiently the mitochondrial dysfunctions over age and figure more accurate the potential safe boundaries of healthy mitochondrial population against the early diagnosis of neurogenerative diseases. Genes, proteins interactions, and mtDNA quality, shall be taken into consideration in future work as the main factors of the proposed representation affecting the mitochondrial population and its functionality.

4. Conclusions and Future Work

Cultural Algorithms will definitely offer a powerful evolutionary algorithmic tool for many biological diseases at both the levels of diagnosis and treatment. In the case of mitochondrial populations and their association with several neurogenerative disorders, this computation technique will lead us to a more formalistic representation of experimental results concerning subcellular measurements and mitochondrial population. In this paper we presented a new combined version of Cultural Algorithms, using the basic notions of evolutionary algorithms on mitochondrial dynamics.

Future research includes more computational tests and a detailed analysis of the performance on real biological data and case studies, concerning mammalian cells.

References

[1] K. J. Batenburg, "An evolutionary algorithm for discrete tomography," *Discrete Applied Mathematics*, vol. 151, no. 1-3, pp. 36–54, 2005.

[2] K. J. Batenburg and W. J. Palenstijn, "A new exam timetabling algorithm," in *Proceedings of the 15th Belgium-Netherlands Artificial Intelligence Conference (BNAIC'03)*, T. Heskes, P. Lucas, L. Vuurpijl, and W. Wiegerinck, Eds., pp. 19–26, 2003.

[3] R. G. Reynolds, "An introduction to cultural algorithms," in *Proceedings of the 3rd Annual Conference on Evolutionary Programming*, pp. 131–139, 1994.

[4] D. E. Goldberg, *Genetic Algorithms*, Addison-Wesley Longman, USA, 1998.

[5] R. G. Reynolds, "An introdution to cultural algorithms," Cultural Algorithms Repository, 1998.

[6] R. G. Reynolds, E. Zannoni, and R. M. Posner, "Learning to understand software using cultural algorithms," Cultural Algorithms Repository, 1998.

[7] A. S. Reichert and W. Neupert, "Mitochondriomics or what makes us breathe," *Trends in Genetics*, vol. 20, no. 11, pp. 555–562, 2004.

[8] R. Lill and G. Kispal, "Maturation of cellular Fe-S proteins: an essential function of mitochondria," *Trends in Biochemical Sciences*, vol. 25, no. 8, pp. 352–356, 2000.

[9] G. Vandecasteele, G. Szabadkai, and R. Rizzuto, "Mitochondrial calcium homeostasis: mechanisms and molecules," *IUBMB Life*, vol. 52, no. 3-5, pp. 213–219, 2002.

[10] A. Trifunovic, "Mitochondrial DNA and ageing," *Biochimica et Biophysica Acta*, vol. 1757, no. 5-6, pp. 611–617, 2006.

[11] D. C. Wallace, "Mitochondrial diseases in man and mouse," *Science*, vol. 283, no. 5407, pp. 1482–1488, 1999.

[12] C. A. Mannella, "The relevance of mitochondrial membrane topology to mitochondrial function," *Biochimica et Biophysica Acta*, vol. 1762, no. 2, pp. 140–147, 2006.

[13] G. Perkins, C. Renken, M. E. Martone, S. J. Young, M. Ellisman, and T. Frey, "Electron tomography of neuronal mitochondria: three-dimensional structure and organization of cristae and membrane contacts," *Journal of Structural Biology*, vol. 119, no. 3, pp. 260–272, 1997.

[14] E. R. S. Kunji, "The role and structure of mitochondrial carriers," *FEBS Letters*, vol. 564, no. 3, pp. 239–244, 2004.

[15] R. Lill and G. Kispal, "Mitochondrial ABC transporters," *Research in Microbiology*, vol. 152, no. 3-4, pp. 331–340, 2001.

[16] B. O'Rourke, "Mitochondrial ion channels," *Annual Review of Physiology*, vol. 69, pp. 19–49, 2007.

[17] L. Palmieri, F. M. Lasorsa, A. Vozza et al., "Identification and functions of new transporters in yeast mitochondria," *Biochimica et Biophysica Acta*, vol. 1459, no. 2-3, pp. 363–369, 2000.

[18] W. Neupert and J. M. Herrmann, "Translocation of proteins into mitochondria," *Annual Review of Biochemistry*, vol. 76, pp. 723–749, 2007.

[19] D. C. Chan, "Mitochondrial fusion and fission in mammals," *Annual Review of Cell and Developmental Biology*, vol. 22, pp. 79–99, 2006.

[20] D. C. Chan, "Mitochondrial dynamics in disease," *New England Journal of Medicine*, vol. 356, no. 17, pp. 1707–1709, 2007.

[21] H. Chen and D. C. Chan, "Mitochondrial dynamics—fusion, fission, movement, and mitophagy—in neurodegenerative diseases," *Human Molecular Genetics*, vol. 18, no. 2, pp. R169–176, 2009.

[22] G. Twig, A. Elorza, A. J. A. Molina et al., "Fission and selective fusion govern mitochondrial segregation and elimination by autophagy," *EMBO Journal*, vol. 27, no. 2, pp. 433–446, 2008.

[23] A. Alexiou, J. Rekkas, and P. Vlamos, "Modeling the mitochondrial dysfunction in neurogenerative diseases due to high H+ concentration," *Bioinformation*, vol. 6, no. 5, pp. 173–175, 2011.

[24] L. J. Martin, "Mitochondrial and cell death mechanisms in neurodegenerative diseases," *Pharmaceuticals*, vol. 3, pp. 839–915, 2010.

[25] R. J. Youle and M. Karbowski, "Mitochondrial fission in apoptosis," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 8, pp. 657–663, 2005.

[26] P. J. Hollenbeck and W. M. Saxton, "The axonal transport of mitochondria," *Journal of Cell Science*, vol. 118, no. 23, pp. 5411–5419, 2005.

[27] S. Campello, R. A. Lacalle, M. Bettella, S. Mañes, L. Scorrano, and A. Viola, "Orchestration of lymphocyte chemotaxis by mitochondrial dynamics," *Journal of Experimental Medicine*, vol. 203, no. 13, pp. 2879–2886, 2006.

[28] H. Chen, J. M. McCaffery, and D. C. Chan, "Mitochondrial fusion protects against neurodegeneration in the cerebellum," *Cell*, vol. 130, no. 3, pp. 548–562, 2007.

[29] D. J. Klionsky and S. D. Emr, "Autophagy as a regulated pathway of cellular degradation," *Science*, vol. 290, no. 5497, pp. 1717–1721, 2000.

[30] A. T. Alexiou, P. M. Vlamos, and K. G. Volikas, "A theoretical artificial approach on reducing mitochondrial abnormalities in Alzheimer's disease," in *Proceedings of the 10th International Conference on Information Technology and Applications in Biomedicine: Emerging Technologies for Patient Specific Healthcare (ITAB'10)*, Corfu, Greece, November 2010.

[31] Z. S. Khachaturian, "Diagnosis of Alzheimer's disease," *Archives of Neurology*, vol. 42, no. 11, pp. 1097–1105, 1985.

[32] S. S. Mirra, A. Heyman, D. McKeel et al., "The Consortium to establish a registry for Alzheimer's disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease," *Neurology*, vol. 41, no. 4, pp. 479–486, 1991.

[33] K. Hirai, G. Aliev, A. Nunomura et al., "Mitochondrial abnormalities in Alzheimer's disease," *Journal of Neuroscience*, vol. 21, no. 9, pp. 3017–3023, 2001.

[34] R. H. Swerdlow and S. M. Khan, "A "mitochondrial cascade hypothesis" for sporadic Alzheimer's disease," *Medical Hypotheses*, vol. 63, no. 1, pp. 8–20, 2004.

[35] R. H. Swerdlow and S. M. Khan, "The Alzheimer's disease mitochondrial cascade hypothesis: an update," *Experimental Neurology*, vol. 218, no. 2, pp. 308–315, 2009.

[36] X. Wang, B. Su, H. Fujioka, and X. Zhu, "Dynamin-like protein 1 reduction underlies mitochondrial morphology and distribution abnormalities in fibroblasts from sporadic Alzheimer's disease patients," *American Journal of Pathology*, vol. 173, no. 2, pp. 470–482, 2008.

[37] C. Geert, *A Interpretação Das Culturas*, Editora Guanabara, Rio de Janeiro, Brazil, 1989.

[38] N. Gessler, *Artificial Culture—Experiments in Synthetic Anthropology*, 1999.

[39] J. Holland, *Adaptation in Natural and Artificial Systems*, MIT Press, Cambridge, Mass, USA, 1992.

[40] J. Barkow, L. Cosmides, and J. Tooby, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford University Press, USA, 1992.

[41] A. Johnson and T. Earle, *The Evolution of Human Societies: from Foraging Group to Agrarian State*, Stanford Univ Press, 2000.

[42] P. Richerson and R. Boyd, *Not by Genes Alone: How Culture Transformed Human Evolution*, University of Chicago Press, 2005.

[43] H. Chen, S. A. Detmer, A. J. Ewald, E. E. Griffin, S. E. Fraser, and D. C. Chan, "Mitofusins Mfn1 and Mfn2 coordinately regulate mitochondrial fusion and are essential for embryonic development," *Journal of Cell Biology*, vol. 160, no. 2, pp. 189–200, 2003.

[44] M. Dobrzyński, J. V. Rodríguez, J. A. Kaandorp, and J. G. Blom, "Computational methods for diffusion-influenced biochemical reactions," *Bioinformatics*, vol. 23, no. 15, pp. 1969–1977, 2007.

[45] M. Długosz and J. Trylska, "Diffusion in crowded biological environments: applications of Brownian dynamics," *BMC Biophysics*, vol. 4, no. 1, Article no. 3, 2011.

[46] D. L. Ermak and J. A. McCammon, "Brownian dynamics with hydrodynamic interactions," *The Journal of Chemical Physics*, vol. 69, no. 4, pp. 1352–1360, 1978.

[47] S. H. Northrup and H. P. Erickson, "Kinetics of protein-protein association explained by Brownian dynamics computer simulation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 8, pp. 3338–3342, 1992.

[48] H. Kim and K. J. Shin, "Exact solution of the reversible diffusion-influenced reaction for an isolated pair in three dimensions," *Physical Review Letters*, vol. 82, no. 7, pp. 1578–1581, 1999.

[49] Z. Frazier and F. Alber, "A computational approach to increase time scales in Brownian dynamics-based reaction-diffusion modeling," *Journal of Computational Biology*, vol. 19, no. 6, pp. 606–618, 2012.

[50] S. Luke, C. Cioffi-Revilla, L. Panait, K. Sullivan, and G. Balan, "MASON: a multiagent simulation environment," *Simulation*, vol. 81, no. 7, pp. 517–527, 2005.

[51] X. Liu, D. Weaver, O. Shirihai, and G. Hajnóczky, "Mitochondrial kiss-and-run: interplay between mitochondrial motility and fusion-fission dynamics," *EMBO Journal*, vol. 28, no. 20, pp. 3074–3089, 2009.