ConsDiff: an algorithm for the detection of conserved differences between protein sequences

Saumil Mehta, Deendayal Dinakarpandian *

Division of Computer Science and Electrical Engineering, School of Computing and Engineering, University of Missouri, 5110 Rockhill Road RHFH550E, Kansas City, MO 64110, USA

Received 22 June 2004; accepted 22 June 2004
Available online 3 August 2004

Abstract

Proteins have been classified into families based on metrics of similarity such as sequence or structural similarity. However, there are significant differences in function even within families. Mapping these differences to individual amino-acid residues is typically done by an expert. This is a subjective and non-scalable approach. ConsDiff is an algorithm that automates this process. It is based on a set of parametric rules using amino-acid substitution matrices and a multiple sequence alignment. This allows the automated discovery of candidate residues that may be responsible for critical differences in function, which may then be experimentally verified.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Algorithm; Knowledge discovery; Multiple sequence alignment; Protein function; Amino-acid substitution matrix; Sequence analysis

* Corresponding author. Tel.: +1 816 235 5942; fax: +1 816 235 5159.
E-mail address: dinakard@umkc.edu (D. Dinakarpandian).

0169-023X/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2004.06.012
1. Introduction

1.1. Organization of paper

We begin with a brief background on bioinformatics that sets the context for the theme of this paper. We then present key concepts that represent the underpinnings of this paper. This is followed by a detailed description of the problem that we address. We then present the proposed solution, describing the algorithm and the implementation with examples drawn from the real world.

1.2. Background

The work described here addresses the general issues of the synergism between bioinformatics and experimental biology, increasing reliance on the comparison of sets of entities, the use of conserved patterns to suggest functional importance, and the need to discover and implement algorithms that can simulate domain experts. We explain each of these general concepts in this section before giving a detailed presentation of the problem in the subsequent section.

1.2.1. Synergism between bioinformatics and experimental biology

Bioinformatics and experimental biology represent a mutually reinforcing relationship where knowledge from one drives the design and development of the other. For example, an important role of bioinformatics today is that of predicting experimental outcomes. In turn, the development of experimental techniques for the large-scale generation of data has spurred research and applications of many areas of computer science like data warehousing, data mining, probabilistic searches and knowledge discovery. It is important to note that, in a large proportion of cases, a prediction made by bioinformatics does not necessarily give a definitive answer. Rather, it merely, but importantly, narrows the search space for verification by subsequent trial-and-error experiments. In this context, too, it serves a very useful purpose, as not all experimental approaches are equally scalable. Many kinds of experimental techniques continue to be highly labor intensive and any computational prediction of limiting the number and/or type of experiments to be performed remains quite valuable. For example, while it is possible to automate the generation of copies of a gene, actual preparation and obtaining pure amounts of the corresponding protein is partially an art that continues to require highly skilled manual intervention.

1.2.2. From one-at-a-time to many-at-a-time paradigm. Corollary: From 1:1 to n:m comparisons

A general class of problems that is important in bioinformatics is that of comparing two objects using some metric of similarity. Depending on the kind of data and level of abstraction, the object may be atomic, a set, a permutation, or a bag (multi-set) of attributes. In biological terms, respective examples are an amino-acid, the proteome, a protein and the genotype (including duplicate alleles).

For example, the score in an amino-acid substitution matrix may be approximated to be an index of similarity between any two amino-acids, normalized to the random evolutionary interchangeability between pairs of amino-acids [10]. Alternatively, at a higher level of abstraction, two proteomes may be subjected to an exhaustive comparison of \( n \) versus \( m \) proteins using a heuristic pair-wise sequence alignment algorithm like BLAST [2]. A composite score of similarity between the two proteomes may be derived from the matrix of pair-wise scores.
1.2.3. Need to encode expert knowledge

The automation of genome sequencing and the use of sequence comparison algorithms have led to the clustering of all known genes. Self-consistent classification schemes like the Clusters of Orthologous Groups (COGS) database [25] further help in distinguishing orthologs from paralogs. In a sense, comparative methods have allowed us to answer the question of “What does each gene/protein do?” However, there are important differences between orthologs from closely related species and even between the variant forms of proteins found in healthy individuals. While global comparison and classification schemes offer useful evolutionary and system-level insights, it is equally important to understand the detailed mechanistic basis of biology—“How does each gene/protein do it?” One of the goals of bioinformatics is the encoding of expert knowledge into software applications. The relatively simple goal of incorporating rules for repetitive tasks is in itself of great value, especially with respect to large volumes of data. Ideally, one would like to incorporate complex decision-making ability based on expert domain knowledge. In this sense, while genome-wide studies are fascinating and offer valuable insights, it is also important to develop algorithms that analyze smaller sets of sequences on the basis of expert knowledge.

1.2.4. Function prediction based on conserved patterns

Given that all living forms share an evolutionary relationship, comparison of proteins with similar function allows the detection of conserved patterns (either subsequences and/or structural motifs) that are likely to have functional significance. The occurrence of a conserved pattern may be due to one of two possible reasons. One is that of functional importance as mentioned in Section 1.2.3. An additional reason, especially in comparing orthologs from closely related species, is simply the lack of time for evolutionary divergence. For example, some human and mouse orthologs are likely to have a very high percentage of identity. However, most of the amino-acid residues that are identical between the two proteins are not likely to be critical for functionality.

In general, the set operations of INTERSECT and DIFFERENCE on the attributes of two biological objects \( i \) and \( j \) will both yield non-empty sets. We relax the requirement for identity to be the similarity above a specified cut-off, in line with the stochastic nature of biology. Let the result of the INTERSECT operation be \( I \) and that of the DIFFERENCE operations be \( D_{i-j} \) and \( D_{j-i} \). Then \( I \) is interpreted to be the set of conserved attributes between the two objects, while \( D_{i-j} \) and \( D_{j-i} \) are the sets of attributes that are unique to objects \( i \) and \( j \) respectively. If the two objects represent instances of data for a certain kind of cancer (with the assumption that the data is relevant to cancer), some subset of \( I \) represents the minimal set of attributes that are associated with the cancer. The limitation of this simple approach, however, is that \( I \) is likely to be much larger than the minimal subset.

2. Problem description

2.1. Motivation

2.1.1. Residues of a protein differ in their weight of contribution to its properties

Proteins have been classified into families on the basis of various approaches based on some metric of sequence [4,24] or structural [7,13,15,19] similarity. However, even with the exclusion
of mutant variants, members of a protein family often exhibit important differences in their properties. For example, small differences at active-site positions between two sequences can be responsible for large changes in functional properties like binding affinities or specificities. These changes need not be restricted to residues present at the active site. For instance, amino-acid substitutions in the core of a protein can disrupt the structure of a protein by altering its folding. At the same time, large differences in positions on the surface of a protein may have negligible effect on protein properties and are readily tolerated. In effect, amino-acid residues that constitute a protein have different weights of contribution to the properties of a protein.

2.1.2. Identifying important residues is crucial to the understanding of function

An important and small subset of amino-acids, \( S_c \), is that of critically important residues whose substitution essentially nullifies the function of the protein. A larger subset \( S_f \) consists of those residues, which have a significant effect on the function of the protein, to varying degrees. For a protein \( p \) consisting of residues \( a_{ij} \) where \( i \) refers to the position in the sequence and \( j \) refers to the amino-acid, the contribution of each residue to the function \( F \) of a protein may be represented as \( w_{ij} \). This has a value of 0 for replaceable residues (in SET operatic terms \( (p - (S_f \cup S_c)) \)) and 1 for critically important ones \( (S_c) \). Values for \( S_f \) would lie between these extremes. Keeping in mind that many proteins have multiple roles, and more than one protein may carry out the same role, it is important to note that the cardinality of mapping between \( n \) proteins and \( m \) roles is \( n:m \). It is important to find \( S_c \) and \( S_f \) in order to understand the details of how a protein works. For instance, protein engineering can be used to alter the properties of a protein by selective alteration (site-directed mutagenesis) of residues that constitute \( S_f \).

2.2. Current approaches to finding important subsets of amino-acids in a protein

2.2.1. 3D Structure-based inference

The availability of the X-ray crystallographic structure of a complex that represents the time-averaged snapshot of the protein in action is very useful in identifying \( S_c \). \( S_c \) may also be inferred from the crystal or NMR structure of the protein. In this case, similarity of the structure \([13,19,22,27]\) or that of the underlying sequence \([2]\) to a well studied example may be used to deduce \( S_c \). \( S_f \) is more challenging to deduce, even if the three-dimensional structure of the protein of interest is already known. This is because dynamic information about the nature of interaction between the protein and the molecules it associates with is missing from the crystal structure.

2.2.2. Inference from mutagenesis

Another line of experiments that is quite important, especially when the 3D structure is unknown, is that of changing \( n \) residues at a time to a different type and measuring the functional behavior of the resulting protein. At one extreme, when nothing is known about which residues might be important, a random mutagenesis approach is used wherein experimental conditions are modified to incorporate a small rate of error into the production of protein. This results in a population of many protein variants, many of which have small differences from the normal (wild type) version. Screening techniques can then be used to identify versions with interesting properties, followed by the actual sequencing of the proteins. At the other extreme, a specific version of a
protein may be engineered with deliberate change at a given position or subsequence and then studied for any changes in a protein property. The latter strategy relies heavily on predictions based on comprehensive and detailed domain knowledge of the protein in question. One approach that lies in between the two extremes is called alanine scanning [9], where the amino-acid alanine is substituted at each of several successive positions in a local region to estimate the importance of each amino-acid that is replaced.

2.2.3. Inference from bioinformatic approaches

Pair-wise local sequence alignment [2,17,21,23] against an annotated database of sequences [1,3,6] based on a sound statistical model is a commonly used approach to infer the functional role of a protein. However, pair-wise sequence analysis has limited value in predicting $S_c$ and $S_f$ as it gives equal importance to matches at all positions.

An alternative approach is to use multiple sequence analysis [12,18,20,26]. The key difference is that the scoring systems used in multiple sequence alignment (MSA) factor in the relative frequency with which pair-wise equivalencies are found. Typically, a researcher scans the output of a multiple sequence alignment program by eye to find the conserved residues in a family. While the percentage of identity or similarity of residues in a multiple sequence alignment of a family is typically quite high, this is in large part due to the lack of sufficient evolutionary time for divergence of the sequences. In other words, the actual residues responsible for the molecular function or some property of the family constitute only a small subset of all conserved residues. One approach to narrowing down the important residues among the conserved residues is to include sequences in the multiple sequence alignment that lack the property of interest. However, for maximum information gain (smaller size of $S_f$), these additional sequences should be similar to the set of sequences being studied. This is analogous to the use of test cases in software engineering that are just valid or just invalid with respect to a boundary condition. In the present context, discovery of the boundary is the goal. Thus, members of a protein family that show internal differences within them with respect to a given property would form such a set. Alternatively, mutants of a protein (see Section 2.2.2) are frequently generated to delineate the contribution of each amino-acid, and multiple sequence alignment may be used to deduce small supersets of $S_f$.

2.3. Problem statement

The problem may be stated as follows:

Consider a multiple sequence alignment of a set $C$ of closely related protein sequences with true subsets $A$ and $B$ that differ in a property. Consider residues $a_i$ ($a_i$ is conserved in $A$, but not in $B$) and $b_j$ ($b_j$ is conserved in $B$, but not in $A$). All $a_i$ and $b_j$ are conserved differences and are candidate residues responsible for the difference in the property of interest.

If $A$ possesses a property that is absent in $B$, all $a_i$ are candidate residues responsible for gain of function, and all $b_j$ are candidate residues responsible for suppression of function. The converse is true as well. The goal of this paper is to present and implement an algorithm that finds and highlights all $a_i$ and $b_j$.

Thus, the problem is one of finding $a_i$ and $b_j$. 
For example, proteins in set $A$ could possess enzymatic activity against a particular substrate while those in set $B$ lack it. Alternatively, members of $A$ and $B$ could both be enzymatically active, but with differing substrate specificity. Gain-of-function refers to the conventional view of a residue contributing to a functional property of the protein. Suppression-of-function implies a blocking action by the residue in question. In other words, the presence of $b_i$ prevents or suppresses the contribution of some residue $x_i$ to the function of the protein; only in the absence of $b_i$ is the residue $x_i$ able to affect a functional property of the protein. This implies that $A$ and $B$ have to be treated symmetrically, i.e., we cannot bias our choice towards either gain-of-function or suppression-of-function.

Note that the term 'subsets' in the statement of the problem refers to subsets of the family of protein sequences, as distinct from residue subsets of a protein. In other words each entity in $A$, $B$ or $C$ is a complete protein sequence and $S_f$ is $\{a_i \cup b_i\}$.

3. ConsDiff

3.1. Overview

The ideal solution should be able to identify $S_f$ based on logic similar to that of a domain expert, in an objective manner, and with different thresholds of confidence within a probabilistic framework.

Given two sets of sequences, ConsDiff (Conserved Differences) uses multiple sequence alignment and parametric rules to find positions conserved in only one of the two sets. The following steps are executed:

1. Two sets of sequences that differ in the property of interest constitute the input.
2. The sequences are first pooled together and multiply aligned using a heuristic multiple sequence alignment algorithm [26].
3. Following this, the sequences are reclassified into the original groups specified by the user.
4. Each column of the multiple sequence alignment is checked for conserved differences. An amino-acid substitution matrix is used along with the ConsDiff algorithm (Section 3.2; Fig. 1) to detect each conserved difference.

3.2. ConsDiff algorithm

The assessment of conserved sequence differences is based on the following algorithm.

1. Minimum internal score (MIS): The MIS at a given position of a subset of multiply aligned sequences is defined as the lowest amino-acid substitution score among all possible pairs of amino-acid residues within the set. This is an index of the lower bound of conservation within each set.
II. Maximum external score (MES): The MES at a given position of 2 subsets of multiply aligned sequences is defined as the highest amino-acid substitution score among all possible pairs consisting of one amino-acid residue from each subset. This is an index of the upper bound of conservation between the sets.

III. Conserved difference score (CDS): The CDS at a given position of two subsets of multiply aligned sequences is defined as \[ \text{CDS} = \max(\text{MIS}(A), \text{MIS}(B)) - \text{MES} \]. In other words, it is the difference between the higher MIS of the two subsets and the MES. This is therefore proportional to the log-odds of the probability of conservation within the more highly conserved set and the probability of conservation across the two sets.

IV. Sensitivity (S): The Sensitivity is a parameter that determines the minimum value of the CDS that is considered significant. The default value of S is 0, corresponding to a log odds-ratio of 1. Higher values of S imply more stringent results.

V. A given position is considered to have a conserved sequence difference if the CDS at that position is greater than S.

3.3. ConsDiff example

\textit{A priori}, each subset of the multiple sequence alignment, considered separately, may consist of identical residues (homogeneous state), similar or dissimilar residues (heterogeneous state) or include at least one gap (gap state). In terms of this classification, this allows for six types of state combinations of the two subsets of sequences that differ in a property. The ConsDiff algorithm is valid for all such combinations, as shown in the following discussion.

Each column/position in the alignment may be considered to be one of the following cases.

\begin{tabular}{|c|c|c|c|c|c|}
\hline
Set & CASE I & II & III & IV & V & VI \\
\hline
A & DG & GL & - & - & D & \\
& DS & I & - & - & D & \\
& DV & V & V & V & D & \\
B & EP & K & I & I & D & \\
& EP & E & M & - & D & \\
& EP & R & R & R & D & \\
\hline
\end{tabular}

\begin{tabular}{|c|c|c|c|c|c|}
\hline
Scoring scheme & MIS (A) & 6 & -3 & 1 & -50 & -50 & 6 \\
& MIS(B) & 5 & 7 & 0 & -3 & -50 & 6 \\
& MES & 2 & -1 & -2 & 3 & 3 & 6 \\
& CDS & 4 & 8 & 3 & -6 & -53 & 0 \\
& Verdict & Y & Y & Y & N & N & N \\
\hline
\end{tabular}

Fig. 1. ConsDiff scoring scheme based on the BLOSUM62 matrix. The final verdict is based on a Sensitivity value of 0 in this example.
Case II: A xor B is homogeneous. In other words, only one of the two sets is homogeneous. For example, B has all Ps at a position while A consists of G, S and V.

Case III: Neither A nor B is homogeneous. In other words, both sets are heterogeneous. For example, A consists of L, I and V while B consists of K, E, and R.

Case IV: A xor B has a gap in at least one of its sequences.

Case V: A and B have a gap in at least one of their sequences.

Case VI: A and B are both homogeneous. For example, all sequences in A and B have aspartic acid (symbol D).

Consider Case I shown in Fig. 1. The BLOSUM62 values for D–D and E–E are 6 and 5 respectively. Since both sets A and B are homogeneous, this corresponds to the MIS values. The MES is 2, the BLOSUM62 value for D–E. The CDS is \([\text{max}(6,5) - 2] = 4\). Assuming a threshold (S) value of 0 indicates that this is a conserved difference. Consider Case II. For set A, the MIS of \( -2 \) is the BLOSUM62 value for G–V, the lowest value among the scores for G–S, G–V and S–V. MIS(B) is 7, the BLOSUM62 score for P–P. The MES is \(-1\), the matrix value for P–S. The CDS is therefore \(7 - (-1) = 8\) and indicates a conserved difference at this position. The scores are similarly derived for Case III. The value of \(-50\) for the MIS in Cases IV and V is the default negative value that is assigned when a gap is compared with an amino-acid.

Intuitively, the CDS derived for the cases shown makes sense. Though D and E share evolutionary similarity (Case I), the fact that there is 100% conservation within each subset that differs in some property suggests that the difference might be important. Case II shows that P is 100% conserved within set B, and the negative value of MES indicates the dissimilarity between P and all members of set A. Case III has set A consisting of members of the equivalence class of hydrophobic amino-acids while set B consists of charged amino acids. Thus, even though all the amino-acids are different in Case III, a conserved difference is apparent. The CDS is negative for Case IV because, even though only set A has a gap in it. Set B does not have conservation among its sequences. Case V is trivial as both subsets have at least one gap, suggesting that this position is inconsequential.

3.4. Algorithmic complexity

Let L be the length of the alignment and n and m be the number of sequences in each set. Then the number of steps required to calculate the two values of MIS at each position is \((\frac{n}{2}) + (\frac{m}{2}) - 1 + (\frac{m}{2}) - 1\). The number of steps required to calculate the MES is \((nm + nm - 1)\). 2 operations are required to calculate the CDS and one comparison with the Sensitivity is required to decide if the value if this is a position of conserved difference. Thus, the complexity of the algorithm is \(2L \left[ \left( \frac{n}{2} \right) + \left( \frac{m}{2} \right) + nm \right] = O(m^2) \sim O(n^2)\) for \(n \sim m\) and linear with respect to \(L\). In other words, the complexity of ConsDiff is negligible compared to that of ClustalW, which is \(O((n + m)^2)\). In practice, a multiple sequence algorithm like ClustalW can align \(\sim10^5\) sequences in a few minutes. For typical datasizes on the order of \(10^4\) sequences, results from ConsDiff are therefore practically instantaneous.
3.5. ConsDiff implementation

Fig. 2 shows the web interface for ConsDiff. ConsDiff has been implemented as a CGI program that may be accessed at http://sice504.ddns.umkc.edu:8888/consdiff.html. Optional features of ConsDiff include a user defined choice of BLOSUM matrix. In addition, the user may upload a custom score matrix. The sensitivity threshold of ConsDiff can be varied from fine to coarse granularity in the definition of equivalence classes, leading to varying levels of stringency in identifying candidate residues.

4. Discussion

4.1. Validation—identification of determinants of collagenolytic ability

Fig. 3 shows a portion of sample output. In this case, a set of collagenases was compared to a set of stromelysins. These are subsets of the family of matrix metalloproteinase enzymes. On the average, collagenases and stromelysins share about 50% sequence identity and have highly similar three-dimensional structures \([5,14,27]\) and active sites. However, only the former have the ability to cleave native Type I collagen \([16]\). In the figure, the top three sequences are stromelysins (set A) while the lower three are collagenases (set B). Examples of Cases III, II and I can be seen at sequence alignment positions 219, 221 and 225 respectively in Fig. 3. Thus, these three positions and others (not shown) are likely candidates for the difference in enzymatic specificity between stromelysins and collagenases. However, our confidence at position 219 is not high because the conserved difference score is 1 — only one above the default Sensitivity of 0. In fact, it has been experimentally shown that partial collagenolytic enzymatic activity can be transferred to...
stromelysins by substituting the subsequence RWTNNFREY from MMP-1 (one of the collagenases in set B) into MMP-3 (one of the stromelysins in set A) [8].

However, this was discovered after extensive trial and error with labor intensive experiments that involved the creation of numerous mutated forms of the proteins. ConsDiff shows that the key residues are only the second N and Y in the subsequence RWTNNFREY (see second sequence from the bottom in Fig. 3), i.e., these constitute $S_f$. Further note that there are no conserved differences between positions 233 to 240. This region in fact constitutes part of $S_c$ (seen readily in the 3D structures that are available in the Protein Data Bank [27]). Hence, as discussed in Sections 2.1.2 and 2.2.1, this is an example of a case where knowledge of the 3D structure is useful, but not sufficient, in understanding the contribution of constituent amino-acids to the function of a protein.

4.2. Justification of scoring scheme

A priori, one might consider some form of average score like the commonly used sum-of-pairs scoring scheme used in multiple sequence alignments. However, choosing maximums and minimums is the appropriate choice, because a lack of conservation is readily proven by negation. For example, even if the average similarity at a given position of a multiple sequence alignment between two sets of sequences is very low, occurrence of the same amino-acid residue in just one instance of each set is sufficient to result in a low CDS.

4.3. ConsDiff versus profile-based alignments

The essential goal of ConsDiff may be viewed as the reduction of search space to a few residues from a larger number of possibilities.

Numerous profile-based approaches such as profile-Hidden Markov Models (HMM) [24] and PSI-BLAST [2] have been used to create families of proteins that exhibit a high degree of similarity. However, at a finer level of discernment, these families are heterogeneous. For instance, in a given HMM, this would correspond to different Viterbi paths for subsets of the sequences. The
(COGS) [25] approach takes discernment a step further in proposing a scheme to distinguish between closely related homologs (orthologs) from not so closely related ones (paralogs). ConsDiff is unique in that it addresses the issue of discernment at the level of amino-acids, and not entire sequences or domains. Also, the approach is not limited to an evolutionary model and thus could be used in protein engineering experiments based on structural or other properties that cannot be readily predicted.

Thus, the best of both worlds would be to use multiple sequence alignments derived from the results of these approaches and combine those with the ConsDiff parametric rules.

4.4. Limitations

4.4.1. Accuracy bounded by choice of matrix and MSA algorithm

Ideally, the threshold for declaring a conserved sequence difference should depend on the three-dimensional structure. For instance, small changes in the core are more significant than large differences in surface residues. However, ConsDiff is based only on primary sequence information. Also, it is important to point out that the predictions are completely dependent on the accuracy of the multiple sequence alignment. Hence, we offer the user the option of submitting customized substitution matrices. Further, given the high degree of similarity of sequences where ConsDiff has an impact, different MSA’s are likely to give highly similar results.

4.4.2. Independence assumption

Similar to the premise that amino-acid substitution matrices are built on, ConsDiff assumes that the contribution of each amino-acid to a functional property of a protein is independent of the contributions of other amino-acids. This is not really true as residues interact with each other in a variety of ways in influencing both structure and function of proteins. The use of a dipeptide [11] substitution matrix might help somewhat in dealing with adjacent dependencies but will not address non-contiguous dependencies—for instance when a set of residues separated by variable gaps have a joint effect on a given function.

5. Conclusions

We have presented an algorithm and a prototype implementation for the objective of automated discovery of conserved differences between two sets of protein sequences. This augments available tools and approaches to detecting conserved sequences by focusing on conserved differences between groups of sequences that differ with respect to a certain property. The default matrices imply an evolutionary basis for comparison. But the approach may be generalized with any matrix to look for candidate residues that are correlated with property differences between sets of proteins. For example, matrices based on physico-chemical properties may be used to gain an understanding of mechanistic issues. We have illustrated the utility of the approach with a real world example taken from protein engineering. In the post-genomic era, in addition to the goal of ascribing function to all known proteins, there is also the need to understand the molecular basis of small differences in function. Sequence similarity-based sequence alignment yields the most
bang for the buck, but is not applicable when the similarity between sequences is too low to yield reliable alignment (≤25%). The ideal solution in this case would be to use ConsDiff with multiple sequence alignments based on structure-structure or sequence-structure alignment. While pairwise alignments based on such approaches is promising, the problem of generalizing such algorithms to multiple sequences is a difficult, but not intractable, one.

References


Saumil Mehta is a graduate student in Computer Science. He obtained his undergraduate degree (Bachelor of Information Technology) from the School of Computing and Engineering at the University of Missouri-Kansas City (UMKC) in August 2002. He began his graduate studies and worked with Dr. Deendayal Dinakarpandian for a year on research projects on Bioinformatic sequence analysis and machine learning. He subsequently transferred to the graduate program at Stanford University. He is currently working on problem-solving methods in the context of the Protégé project at Stanford Medical Informatics.

Deendayal Dinakarpandian is currently an Assistant Professor of Computer Science in the School of Computing and Engineering at the University of Missouri-Kansas City (UMKC). He obtained his M.B.B.S. (Bachelor of Medicine/Bachelor of Surgery) and M.D. degrees from the Juhawaral Institute of Post-Graduate Medical Education and Research in Pondicherry, India. He obtained his Ph.D. degree in Biochemistry from Case Western Reserve University in Cleveland, USA, in 1999. He then worked as a post-doctoral researcher at the Kansas University Medical Center (1999–2002) and obtained an M.S. in Computer Science from UMKC in 2002. He was a Visiting Scientist at the Stowers Institute for Medical Research, Kansas City, USA, in summer 2002 before starting at his present position. His research has been in the area of molecular modeling, computational chemistry and Raman spectroscopy. His current research interests include algorithms and machine learning approaches for the analysis of biological data, modeling biological systems and designing analytical interfaces for biological databases.