Global Journal of Computer Science and Technology
Volume 12 Issue 1 Version 1.0 January 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Hotspot Identification System for identification of core residues in Diabetic Proteins

By P.V.S.L. Jagadamba, M.S.Prasadbabu, Allam Apparao
Professor, Dept of CS&SE, Andhra University, Visakhapatnam

Abstract - Data on genome structural and functional features for various organisms are being accumulated and analyzed in laboratories all over the world. The data are stored and analyzed on a large variety of expert systems. The public access to most of these data offers to scientists around the world an unprecedented chance to data mine and explores in depth this extraordinary information repository, trying to convert data into knowledge. The DNA and RNA molecules are symbolic sequences of amino acids in the corresponding proteins has definite advantages in what concerns storage, search, and retrieval of genomic information. In this study an attempt is made to develop an algorithm for aligning multiple DNA / protein sequences. In this process hotspots are located in a protein sequence using the multiple sequence alignment.

Keywords : *Symbolic sequences, DNA, RNA, Protein sequence, Multiple Sequence alignment.*

GJCST Classification: *Optional, DDC/LCC/UDC/Global Journals/NLMC/FOR/MSC Classifications Accepted*

Strictly as per the compliance and regulations of:

Hotspot Identification System for identification of core residues in Diabetic Proteins

P.V.S.L. Jagadamba[α], M.S.Prasadbabu[Ω], Allam Apparao[β]

Abstract - Data on genome structural and functional features for various organisms are being accumulated and analyzed in laboratories all over the world. The data are stored and analyzed on a large variety of expert systems. The public access to most of these data offers to scientists around the world an unprecedented chance to data mine and explores in depth this extraordinary information repository, trying to convert data into knowledge. The DNA and RNA molecules are symbolic sequences of amino acids in the corresponding proteins has definite advantages in what concerns storage, search, and retrieval of genomic information. In this study an attempt is made to develop an algorithm for aligning multiple DNA / protein sequences. In this process hotspots are located in a protein sequence using the multiple sequence alignment

Keywords : Symbolic sequences, DNA, RNA, Protein sequence, Multiple Sequence alignment.

I. Introduction

In Bioinformatics, sequence alignment is a prominent method of arranging the sequences of DNA, RNA or protein to identify regions of similarity. Similarity may be functional, structural or evolutionary relationships between the sequences. Aligned sequences of nucleotide, amino acid residues are represented in a row form of a matrix. Identical or similar characters are aligned in successive columns by inserting gaps between the residues. There is a storm of revolution in the areas of Genomics and Bioinformatics in recent years. Bioinformatics is widely used for computational usage and processing of molecular and genetic data. The biologists considered Bioinformatics for the use of computational methods and tools to handle large amounts of data and make the data more understandable and useful. On the other hand, others view Bioinformatics as an area of developing algorithms and tools and to use mathematical and computational approaches to address theoretical and experimental questions in biology. As genomic data is rapidly exposed to increasing research, knowledge based expert system is becoming indispensable for the emerging studies in Bioinformatics. Hence validation and analysis of mass experimental and predicted data to identify relevant biological patterns and to extract the hidden knowledge are becoming important.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                     ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .****************    :  *.: :
```

In recent years, semantic web based methods are introduced and are designed in such a way that meaning is added to the raw data by using formal descriptions of concepts, terms and relationships encoded within the data. To analyze and understand the data, today's information rich environment developed and designed a number of software tools. These tools provide powerful computational platforms for performing Insilco experiments (8). As there is much complexity and diversity in the analysis of tools, the need is for an intelligent computer system for automated processing. Present researches in Bioinformatics need the use of integrated expert systems to extract more efficient knowledge. In the biological process proteins undergo some interactions. These protein-protein interactions are mediated molecular mechanisms. During this interaction, a small set of residues play a critical role. These residues are called hot spots. The ability to identify the hot spots from sequence accurately and efficiently as expert system that enables and analysis of protein-protein interaction hot spots. This analysis may benefit function prediction and drug development. At present there is a strong need for methods to obtain an accurate description of protein interfaces. Many scientists try to extract protein interaction information from protein data bank.

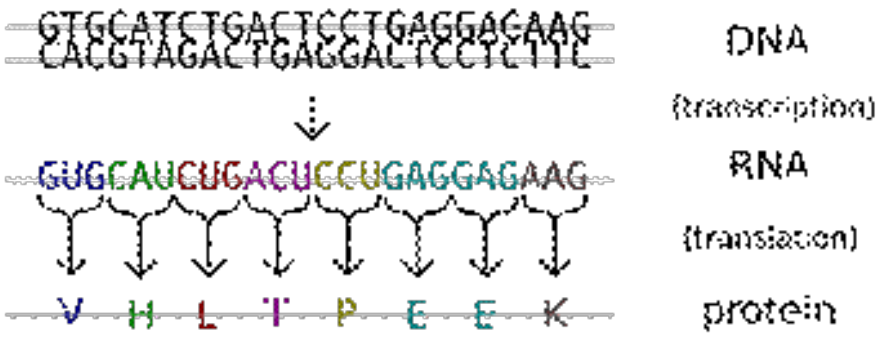
Alignment Methods Used: In general the hot spots are identified as active sites in protein structures as binding is done using structures. The researcher tried

Author α : Principle Investigator, Women Scientist Scheme (WOS-A), DST Project, JNTUK, Kakinada, AP, India.
Author Ω : Professor, Dept of CS&SE, Andhra University, Visakhapatnam
Author β : Vice Chancellor, JNTUK, Kakinada

to find the hotspots in protein sequence rather than structure. In this process, taking into consideration the evolutionary history, the families of sequences are aligned using multiple sequence alignment.

In the process of alignment two methods are used Standard method using dynamic programming and A proposed alternative- MSAPSO (Multiple Sequence alignment using Particle Swarm Optimization) method in which alignment is performed using PSO technique. A comparison of these two methods also made. If the sequences are very short or similar they can be aligned by hand. But lengthy and highly variable numerous sequences cannot be aligned manually. To produce high quality sequence alignments, construction of algorithms and application of human knowledge are necessary. Computational approaches to sequence alignments are of two types- Global alignments and local alignments. Global alignment is the alignment to span the entire length of sequences whereas local alignments identify regions of similarity within the long sequences.

1. Particle Swarm Optimization: Particle Swarm Optimization (PSO) is based on stochastic optimization technique. It is one of the machine learning algorithms. It has been considered to be an effective optimization tool in many areas. The interesting point in PSO is that each particle with potential solution searches through the problem by updating itself with its own memory and also the social information gathers from other particles. Multiple Sequence Alignment: When three or more biological sequences namely protein, DNA or RNA are generally aligned, it is called multiple sequence alignment. As it is difficult and also time consuming to align by hand, computational algorithms are used to analyze and produce such biological sequences. Most multiple sequence alignment programs use heuristic methods as the order of the sequences to align plays a vital role. Development of MSA algorithm is now an active are of research. MSA alignments are an essential tool for protein structure and function prediction, phylogeny inference and other common tasks in sequence analysis.
2. Pair wise Sequence Alignment: If two sequences are arranged for an alignment it is known as pair wise sequence alignment. The degree of relationship between the sequences is predicted computationally or statistically based on weights assigned to the elements aligned between sequences. The standard algorithm to align a pair of sequences is Needleman Wunch algorithm. This algorithm uses dynamic programming. In this study an algorithm PSAPSO (Pair wise Sequence alignment using Particle Swarm Optimization) is proposed and is also compared with the standard algorithm to know the accuracy of the results. A gene encoded in the genetic code defines the amino acid sequence in a protein. An amino acid residue is the combination of three nucleotides. Each three-nucleotide set is a codon. The set of codons forms a genetic code. For example AUG stands for methonine M. In this AUG is a codon, M is an amino acid and the residues A, U, G are nucleotides. Genes encoded in DNA are first transcribed into pre-messenger RNA (mRNA) known as primary transcript. Then pre-mRNA process to mature mRNA using various forms of modifications of posttranscriptional modifications. Then mature mRNA is used as a template for protein synthesis, which is known as translation onto a ribosome. Then read three nucleotides at a time by matching each codon to its base pairing anticodon to form transfer RNA (tRNA). Then tRNA recognizes the amino acid corresponding to the codon. The sequence thus obtained is protein sequence.



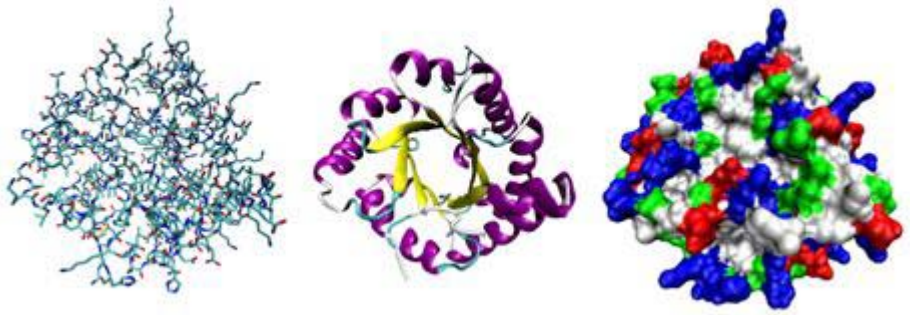
The amino acids in a protein sequence are shown in the following table.

Table 1

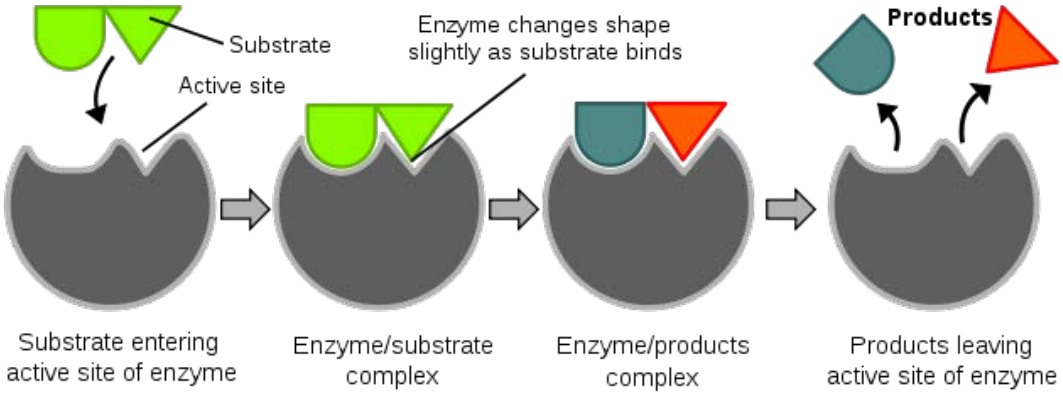
One Letter	Three Letter	Full Name	One Letter	Three Letter	Full Name
G	GLY	Glycine	W	TRP	Tryptopham
A	ALA	Alanine	Y	TYR	Threonine
V	VAL	Valine	N	ASN	Asparagine
L	LEU	Leucine	Q	GLN	Glutamine
I	ILE	Lsoleucnie	D	ASP	Asparatic Acid
F	PHE	Phenylalanine	E	GLU	Glutamic Acid
P	PRO	Proline	K	LYS	Lysine
S	SER	Serine	R	ARG	Arginine
T	THR	Threonine	H	HIS	Histidine
C	CYS	Cyctenie	M	MET	Methinine

The overall structure and function of a protein is determined by the amino sequence. Most proteins fold into 3-dimensional structures and its shape is known as its native state. There are four levels in a protein structure.

- Primary Structure: Primary structure is nothing but an amino acid sequence. Secondary Structure: Secondary structures are regularly repeating local structures and are stabilized by hydrogen bonds. As they are local in nature different secondary structures can be present in the same protein molecule. Example alpha helix, beta sheet and turns.
- Tertiary Structure: Tertiary structure is the special relationship of the secondary structures to one another and is generally stabilized by the formation of the hydrophobic core, a non-local interaction. Salt bridges, hydrogen bonds; disulphide bonds and even post-transnational modifications also stabilize it. It mainly controls the basic function of the protein.
- Quaternary Structure: This structure is formed by several protein molecules i.e. poly peptide chains and it functions as a single protein complex.



- Enzymes: Enzyme is one of the functions of the protein which carries out most of the reactions involved in metabolic activities. Enzymes are proteins that increase the rate of chemical reaction. Adding or participation of the substance called catalyst does the change in the rate of chemical reaction. Catalysts that speed the reaction are called positive catalysts. Substances that interact with catalysts to slow the reaction are called inhibitors (or negative catalysts). Substances that increase the activity of catalysts are called promoters, and substances that deactivate catalysts are called catalytic poisons.
- Active Sites in Proteins: An Active site is a part of an enzyme where substrates bind and undergo a chemical reaction. The substrate which is a molecule binds with the enzyme active site and then an enzyme-substrate complex is formed. It is then transformed into one or more products, which are released from the active site. The active site is now free to accept another substrate molecule. In the case of more than one substrate, these may bind in a particular order to the active site, before reacting together to produce products. A product is something "manufactured" by an enzyme from substrate. For example the products of its Lactase are Galactose and Glucose, which are produced from the substrate Lactose.



Two models- the lock and key model and induced fit model are the two models proposed to describe how the enzymes work. In the lock and key model the active site perfectly fits for a specific substrate. If once the substrate binds to the enzyme no further modification is necessary. On the other hand in the induced fit model, an active site is more flexible and the presence of certain residues (amino acids) of the active site the enzyme is encouraged to locate the correct substrate. Once the substrate is gone conformational changes may occur. Hot spots are a set of residues recognized or bound in the process of interacting with other proteins. These are the residues in the active site.

II. Results & Discussion

Insulin is one of the important protein sequences which cause diabetes. So we tried to identify the hotspots in this protein sequence using the following methodology.

- The protein structures are retrieved from protein data bank by mapping with insulin protein sequence accession p01038 shown in the following table.

SNO	PDB Code	Chain	First PDB residue	Last PDB residue	First P01308 (INS_Human) residue	Last P01308 (INS_Human) residue
1	1a7f	A	1	21	90	110
2	1a7f	B	1	29	25	53
3	1ai0	A	1	21	90	110
4	1ai0	B	1	30	25	53
5	1ai0	C	1	21	90	110
6	1ai0	D	1	30	25	54
7	1ai0	E	1	21	90	110
8	1ai0	F	1	30	25	54
9	1ai0	G	1	21	90	110
10	1ai0	H	1	30	25	54
11	1ai0	I	1	21	90	110
12	1ai0	J	1	30	25	54

12	1ai0	J	1	30	25	54
13	1ai0	K	1	21	90	110
14	1ai0	L	1	30	25	54
15	1aiy	A	1	21	90	110
16	1aiy	B	1	30	25	53
17	1aiy	C	1	21	90	110
18	1aiy	D	1	30	25	54
19	1aiy	E	1	21	90	110
20	1aiy	F	1	30	25	54
21	1aiy	G	1	21	90	110
22	1aiy	H	1	30	25	54
23	1aiy	I	1	21	90	110
24	1aiy	J	1	30	25	54
25	1aiy	K	1	21	90	110

- Then identify the protein-protein interactions for each of these protein structures shown in the following table.

SNO	PDB Code	Chain	Chain
1	1a7f	A	B
2	1ai0	A	B
3	1ai0	B	D
4	1ai0	C	D
5	1ai0	E	F
6	1ai0	F	H
7	1ai0	G	H
8	1ai0	I	J
9	1ai0	J	L
10	1ai0	K	L
11	1aiy	A	B
12	1aiy	B	D
13	1aiy	C	D
14	1aiy	E	F
15	1aiy	F	H
16	1aiy	G	H
17	1aiy	I	J
18	1aiy	J	L
19	1aiy	K	L
20	1b9e	A	B
21	1b9e	B	D
22	1b9e	C	D
23	1guj	A	B
24	1guj	B	D
25	1guj	C	D

Identification of Hotspot: The hot spots are identified using these interfaces and the hot spots in the protein sequence p01308 are

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTR

REAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

III. Conclusion

Hot spots are of residues comprising only a small fraction of interfaces of the binding energy. We present a new and efficient method to determine computational hot spots based on pair wiser technique using potentials and solvent accessibility of interface residues. The conservation does not have significant effect in hot spot prediction as a single feature. Residue occlusions from solvent and pair wise potentials are found to be the main discriminative features in hot spot prediction. The predicted hotspots are observed to match with the experimental hot spots with an accuracy of 70%. The solvent is a necessary factor to define a hot spot, but not sufficient itself. This is also compared our methods and other hot spot prediction methods. Our method outperforms them with its high performance expert system.

References References Referencias

1. Chao-Yie Yng and Shaomeng Wang, "Computational Analysis of Protein Hotspots",ACS Medicinal Chemistry Letters, 2010,1 (3) pp 125-129.
2. Hajduk, P.J et al (2005) "Druggability induces for protein targets derived from NMR-based Screening Data", J Med. Chem. 48, 2518-2525.
3. Dobson CM. (2000). The nature and significance of protein folding. In Mechanisms of Protein Folding 2nd ed. Ed. RH Pain. Frontiers in Molecular Biology series. OxfordUniversity Press: New York, NY.
4. Hintze Miller B.(1988), "Expert System An Introduction" PC AI where Intelligent technology meets the real world, 2(3), 26.
5. Robert S. Engelmopre, Edward Feigenbaum, 1993, "Expert Sstems and Artificial Intelligence", WTEC Hyper Librarian.
6. Mohamed Radhouene Aniba and Julie D. Thompson, "Knowledge Based Expert Systems in Bioinformatics", published in Expert Systems, Book edited by: Petrică Vizureanu, ISBN 978-953-307-032-2, pp. 181-192, 2010.
7. Roos DS. Computational biology. Bioinformatics – trying to swim in a seaof data. Science. 2001;291:1 260—1.