Domain Hierarchy of Protein Loop-Lock Structure (DHoPLLS): a server for decomposition of a protein structure on set of closed loops.

Simon B. Kogan*, Oleg Kupervasser**

Genome Diversity Center, Institute of Evolution, University of Haifa,

Haifa 31905, Israel.

*Email: simonkog@gmail.com

** Email: olegkup@yahoo.com

Abstract

DHoPLLS (http://leah.haifa.ac.il/~skogan/Apache/mydata1/main.html) is a web server that identifies closed loops, which constitute a structural basis for the protein domain hierarchy. The server was created in 2005 year on basis Prof. Trifonov's lab in Genome Diversity Center, Institute of Evolution at University of Haifa. It is based on theory of loop-lock structure developed by Prof. Trifonov's group in Weizmann Institute of Science. This theory declares that all globular protein can be decomposed on set of almost closed loop with mean length about 30 aminoacids. The lock is a place, where a loop meets itself. Investigations of Trifonov's group demonstrates that some aminoacid's sequences (about 30 aminoacids) are usually exists in proteins in form of closed loops. On basis computer programs, developed for the server, we checked the simplest assumption that most of closed loop or locks in proteins have some alphabet of consensus sequences. Unfortunately our results demonstrate that such alphabet doesn't exist. But may be some more complicate algorithm exists for loop finding from aminoacid's sequence of protein.

1. Introduction

Usually for protein 3D structure analysis domains are used. In last years by new challenging method was assume for 3D protein analysis on basis loop-lock structure Prof. Trifonov's group in Weizmann Institute of Science [1-5]. It was proved that any globular protein can be decomposed on set of almost closed loop (with mean length about 30 aminoacids) and locks, which are the nearest not adjoining parts of the closed loops. It was proved that some aminoacid's sequences are usually exists in proteins in form of such closed loops. The future development was made in Prof. Trifonov's group in Haifa University. It is [6-9] or the described web server and the papers, published there. It is, for example, an attempt to get a consensus sequence's alphabet for loops and locks. Results of this attempt (unfortunately unsuccessful) will be presented here. Nice algorithm for loop-lock structure finding was developed and described on the site.

The rest of paper is following. In part 2 we describe possibilities of the DHoPLLS. In part 3 we explain algorithm of decomposition on loops and locks. In part 3 we present results of this attempt to find alphabet for loops and locks. In part 4 we make conclusions.

2. DHoPLLS implementation.

First of all for correct work with DHoPLLS it is necessary to download Chime program for proteins 3D Visualization in Internet Explorer which allows to move them by the help of a computer's mouth. The relevant reference can be found on the site. In the right part of the site we can see "menu". It includes (Fig.1):

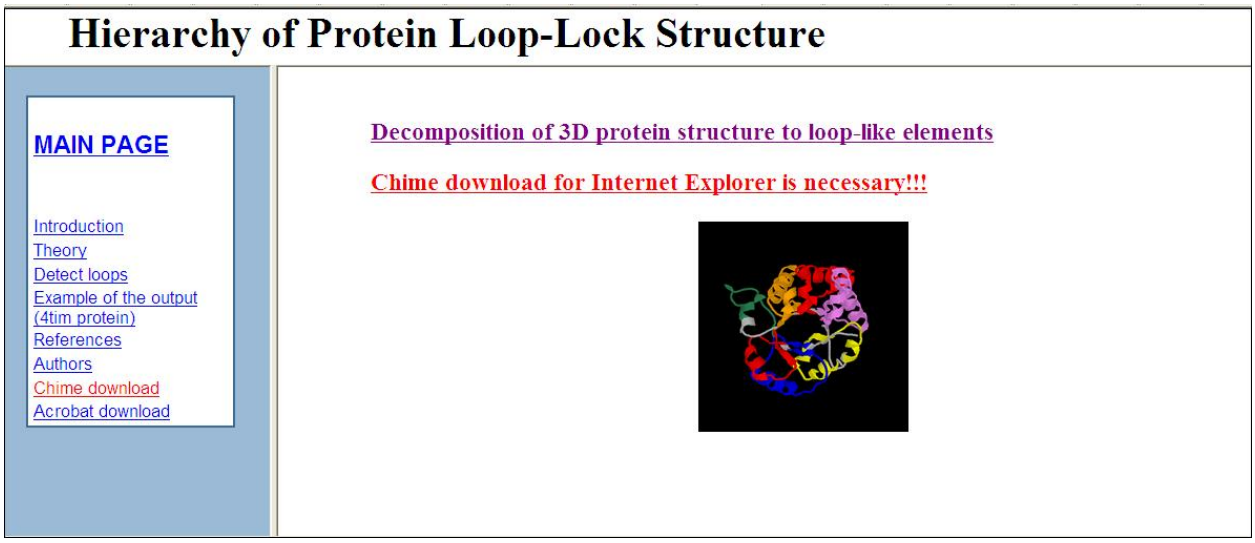

Fig. 1. Initial page of the site with "Menu".

I) Main page which gives references to page for the loop structure calculation and reference to the Chime download site.

II) Introduction – explain algorithm for loop-lock structure finding

III) Theory – it gives the most relevant publication for site topic.

IV) **Detect loop** – it is the main part of site for loop structure calculation.

V) **Example of output** – it is example of resulting page of "Detect loop" job for 4tim-protein.

VI) References – list of relevant papers

VII) Authors – list of references to homepages of the site authors

VIII) Chime download – reference for Chime download

IX) Acrobat download – reference for Acrobat download

Let us described "**Detect loop"** initial page. Three ways to make input of a protein exists (Fig. 2):

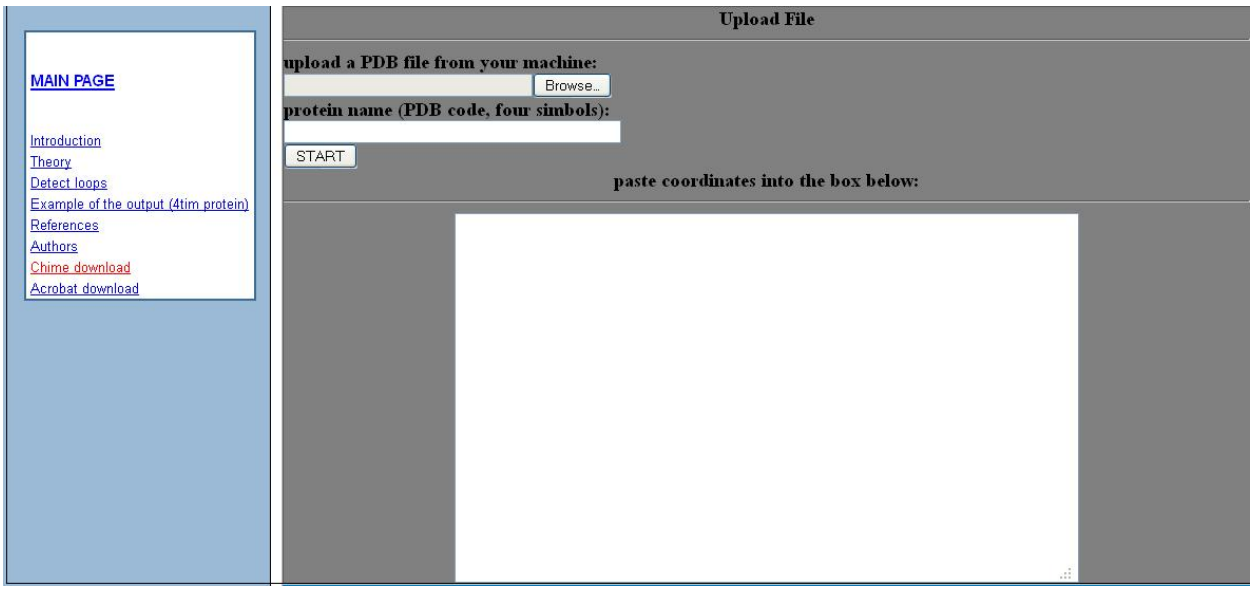

Fig.2 "Detect loop" page for input of a protein and for a loop-lock structure detection.

I) From "pdb" file, existing on user's computer

II) To make input of pdb file inside "window" on the site

III) To give PDB code (4 chars: 4tim, for example). The DHoPLLS downloads himself this protein from relevant Protein Data Bank in Internet.

Example of output page can be found in "**Example of output"** (Fig. 3).

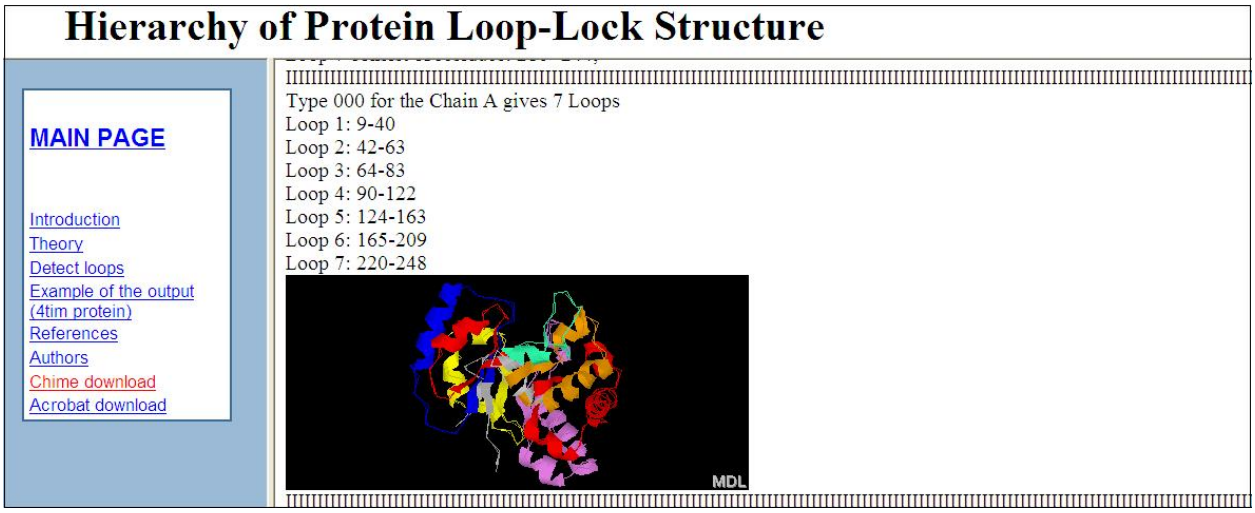

Fig. 3 Output page of algorithm (for loop-lock structure decomposition) operation.

It is a list of the best five decompositions of a protein on loops (for every chain of protein). Indeed, ambiguity in such decomposition exists. Below every list of loops (with the number of the first and the last aminoacid) we can find figure of protein with loops colored to different colors. We can rotate the protein by computer's mouth. It is result of Chime Visualization.

3. Algorithm for decomposing protein structure into loop-like elements

Early evolution of biological molecules is one of the most important questions in evolution of life. This is very controversial and largely unknown area. The loop-and-lock protein structure theory [1,3] intends to close the gap in our understanding of the rules and order of protein domain assembly from individual amino acids. According to the theory the basic element of protein structure is a closed loop with an average size of 30 amino acid residues. The early proteins were those loop-like structures on their own. Then they congregate into larger units, subdomains and domains. The secondary structure elements, -helices and -sheets, play rather auxiliary role in this process. The loop preferential size is calculated on the base of polymer statistics lows and is governed largely by polypeptide chain flexibility. Using the predictions of the theory, several major types of ancestral loop elements were reconstructed on the base of data collected from 23 complete bacterial proteomes [3].

This website was constructed to illustrate some of the software developed in the course of loop and lock protein structure investigation. The input of the program is a protein structure in the pdb format. The output is several variants of the structure decomposition into loop elements.

The algorithm is as follows:

Given three-dimensional protein structure, find all amino acid sequences (within certain length limits) that have ends complying with two conditions: the ends have to be closer to each other than certain threshold; the distance between the ends has to be the local minimum among all possible distances between neighboring to the ends amino acids. The sequences can overlap and in many cases they indeed overlap each other. Each such sequence is a potential protein loop.

Give a weight to each potential loop according to its sequence length and the distance in three-dimensional space between end amino acids. Find maximum weight independent (non-overlapped) set of potential loops. This is the first and the best decomposition variant. To obtain second best decomposition, each loop from the first decomposition is excluded in turn and search for maximum weight independent set among remaining potential loops is performed. The number of these partial decompositions is equal to the number of loops in the first decomposition. The best (having maximum weight) among partial decompositions is chosen.

To obtain third best decomposition, each loop from the second decomposition is excluded in turn and search for maximum weight independent set among remaining potential loops is performed. Then the best decomposition among vacant partial decompositions of the first and second decompositions is chosen.

The following ranked decompositions are obtained similarly. The maximum weight independent set search is done according to the simple and beautiful procedure in linear time [10].

4. Some negative results about possibility of protein code for loops or locks

The “loop and lock” protein evolution theory considers a protein as a sequence of loop like elements (with average length about 30 amino acids). The ends of these loops are referred to as locks [4]. One prediction of the theory is the existence of loop "alphabet", i.e., existence of a finite and small set of amino acid sequences that are ancestors of all modern loop elements. Any modern protein loop should be close (in sequence sense) to one “letter” of this alphabet. Hereafter, two sequences are said to be close if they retain at least 30% similarity. The probability of this to happen randomly is very small and equal approximately to 1/100000 [5]. Another prediction of the theory is the presence of similar "alphabet" for locks. We define two lock sequences to be closely related if these sequences are matched exactly, since locks are very

short. If these two predictions are true, it would allow us to find loops or locks without knowledge of 3D protein structure.

This work has been carried out with the purpose of confirming the above predictions. About 6000 proteins with known 3D structure were examined in the experiment. These proteins have been decomposed into approximately 100000 loops. If "alphabet" of loops actually exists, they can be clustered into a small number of groups in such a way that all loops in the same group are close to each other. Since the probability of random closeness is 1/100000, we expect to form about 50000 groups in the absence of any "alphabet".

For finding of groups the following algorithm has been used:

I) Take one loop from the whole loop set and compare it to all other loops in the set. Form one group from all loops that are close to the first loop.

II) Exclude the group loops from the set.

III) Return to the first step until the set is empty.

The number of groups found in the experiment is about 40000 that is too large to claim the "alphabet" existence. More complex loop comparison model, allowing for insertions or deletions of amino acids has also been considered [11]. The number of groups found this way is about 30000 that is still too large.

The option of a "binary" code [6-7] has been considered as well. Thus, 20 amino acids are clustered into two groups originated by ALA and GLY amino acids. To retain 1/100000 threshold of closeness, the 90% percent loop similarity is demanded in the case of “binary” code. The number of groups found in "binary" code experiment was close to 50000.

Also the hypothesis of lock "alphabet" existence was tested. The lock is defined as the last three or five amino acids on both loop ends. The number of groups found in lock clustering experiments was found to be close to random expectation.

On the basis of this work the following conclusions can be made:

I) The obtained results do not confirm the existence of loop or lock "alphabet". The results do not contradict the fact that there is a small number of amino acid sequences that are usually loops or locks in proteins [4]. Majority of loops and locks, however, have different sequences.

II) Another algorithm (more advanced than simple 30% similarity) could probably solve this problem.

5. Conclusions

On basis the original algorithm for look-lock decomposition, which was developed in this paper, the web server was implemented. This server identifies the closed loops, which constitute a structural basis for the protein domain hierarchy. This web server was made formerly than similar one [12]. On basis developed computer’s program was investigate possibility of the loop-lock’s alphabet existence. Obtained negative result is also important for understanding 3D structure of proteins.

Acknowledgment
We would like to thank Edward N. Trifonov for his supervision and for many fruitful ideas used in this paper. We also would like to thank Zakhar M. Frenkel for many fruitful ideas used in this paper.

References

[1] Berezovsky I.N., Grosberg A.Y., Trifonov E.N. "Closed loops of nearly standard size: common basic element of protein structure", FEBS Lett., **466**, P.283–286. (2000)

[2] Berezovsky I.N., Trifonov E.N. "Van der Waals locks: loop-n-lock structure of globular proteins", J. Mol. Biol., **307**, P.1419–1426 (2001)

[3] Trifonov, E. N., Berezovsky, I. N., "Proteomic code", Molecular Biology, **36**, P.239-243 (2002)

[4] Berezovsky I.N., Kirzhner A., Kirzhner V.M., Trifonov E.N., "An Eye-Opener to Protein Structures", ComPlexUs, 1, P.29–37 (2003)

[5] Berezovsky I.N., Kirzhner A., Kirzhner V.M., Rosenfeld V.R., Trifonov E.N. , "Protein sequences yield a proteomic code", J Biomol Struct Dyn., **21**(3), P.317-325 (2003)

[6] Trifonov E.N., J. Biomolec. Str. Dyn., **22**, P.1-11 (2004).

[7] Trifonov E.N., "Theory of early molecular evolution: Predictions and confirmations": In Book "Discovering Biomolecular Mechanisms with Computational Biology", Ed. F. Eisenhaber, Landes Bioscience, Georgetown, P.107-116. (2006)

[8] Zakhar M. Frenkel Z.M., Trifonov E.N., "Closed Loops of TIM Barrel Protein Fold Journal of Biomolecular Structure and Dynamics", **22** (6), P.643-655 (2005)

[9] Aharonovsky E., Trifonov E.N. "Sequence structure of van der Waals locks in proteins", J Biomol Struct Dyn., **22**(5), P.545-53 (2005)

[10] Hsiao, J. Y., Tang, C. Y. , Chang, R. S. "An Efficient Algorithm for Finding a Maximum Weight 2-Independent Set on Interval-Graphs", Information Processing Letters, **43**, P. 229-235 (1992)

[11] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis, Cambridge University Press, 356 p, (1998)

[12] Koczyk G., Berezovsky I.N., "Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of protein domain structure", Nucleic Acids Research, 36 (Web Server issue), P.239-245 (2008)