БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ • BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, **56**
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, **56**

София • 2006 • Sofia

The Protein Sequence Structure Alignment – Computer's Problem in the Face of Molecular Biology

Nikola Chakarov

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 1164 Sofia
E-mails: nikola.chakarov@gmail.com chakarov@fmi.uni-sofia.bg

1. Introduction

It is well known that the chemical composition of the living organisms is relatively constant. About 70% from every cell is water. About 4% are small molecules. The proteins build about 15% to 20% of the cell; DNA and RNA are 2-7% from cell's weight.

The rest 4-7% are cell's membranes lipids and other molecules.

The problem concerning Sequence Structure Alignment is extremely important, because the biology activities of the proteins are determinate mainly from their three-dimensional folded shape in the space.

2. Structure of proteins

Proteins are long chains of amino acids. There are 20 different amino acids that serve as building blocks for proteins. Amino acids have a specific chemical structure which contains a carbon backbone similar to all amino acids and a residue which varies between the amino acids.

Amino Acids (Fig. 1): Although the well known are only about 20 amino acids, there are about six more found in the body. Many others are also known from a variety of sources. Amino acids are the building blocks used to make peptides and proteins (Fig. 2). The different amino acids have interesting properties because they have a variety of structural parts which result in different polarities and solubilities. Each amino acid has at least one amine and one acid functional group as the name implies. The different properties result from variations in the structures of different R groups. The R group is often referred to as the amino acid "side chain". Amino acids have special common

names, however, a three letter abreviation for the name is used most of the time (see [1]) .

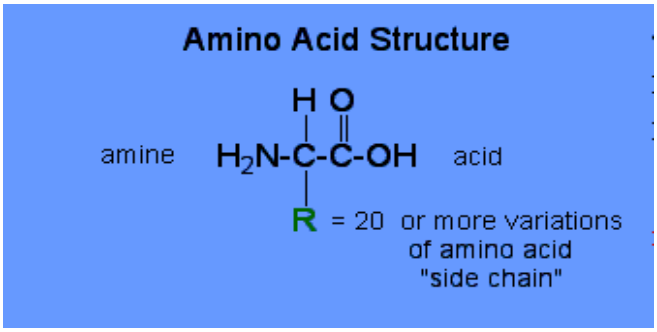

Fig. 1. Amino acid structure

Proteins are the most important class of biochemical molecules, although of course lipids and carbohydrates are also essential for life. Proteins are the basis for the major structural components of animal and human tissue. Proteins are natural polymer molecules consisting of amino acid units. The number of amino acids in proteins may range from two to several thousand.The primary protein structure is defined as the specific sequence of amino acids in the protein.

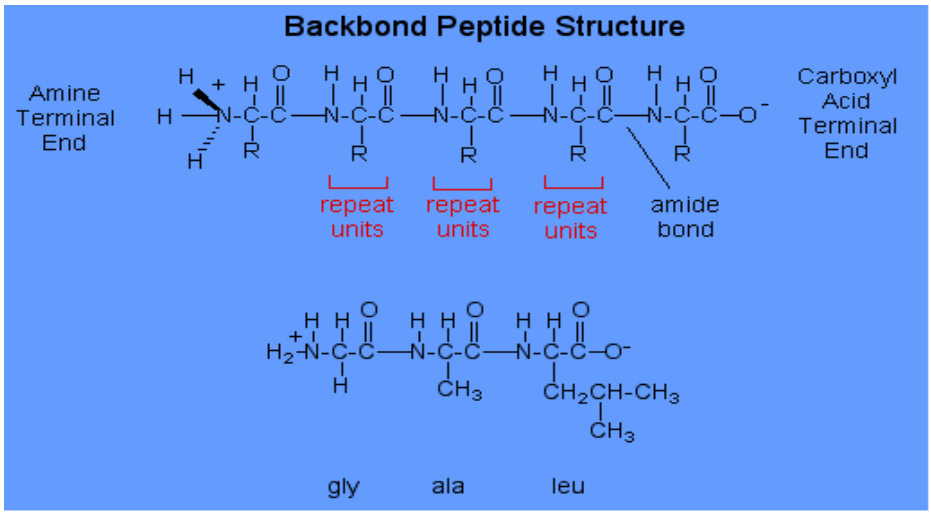

Fig. 2. Backbond peptide structure

In order to function properly, peptides and proteins must have the correct sequence of amino acids [2]. The dipeptide which consisting of two different amino acids could have two different sequences as in the example gly-ala or ala-gly.

It is very important that as written left to right in gly-ala, the glycine has the “free” amine terminal end and alanine has the “free” carboxyl acid terminal end.

If three different amino acids (gly, ala, leu) are used to make a tripeptide, there are six possible sequences:

gly - ala - leu; gly - leu - ala; ala - gly - leu;

ala - leu - gly; leu - ala - gly; leu - gly - ala.

The length of a protein chain can range from 50 to over 3000 amino acids. Proteins are known to have many important functions in the cell, such as enzymatic activity, storage and transport of material, signal transduction, antibodies and more. An important property of a protein is the length and composition of the amino acids chain. The series can be obtained automatically from the gene that encodes for the protein. Another interesting property is the unique folding. The amino acids composition of a protein will usually uniquely determine the 3D structure of the protein. That means two proteins with the same amino acids sequence will have the same 3D structure in natural condition. All proteins whose structure is known are stored in the Protein Data Bank (PDB) [5].

There are multiple levels of structure of the proteins [3] (see Fig. 3):

- Primary (linear) structure – Chain of amino acids.
- Secondary structure – Chains of structural regular elements, most important of which are α-helices and β-sheets.
- Tertiary and Quaternary structure – 3D structure, of a single amino acids chain or several chains, respectively.

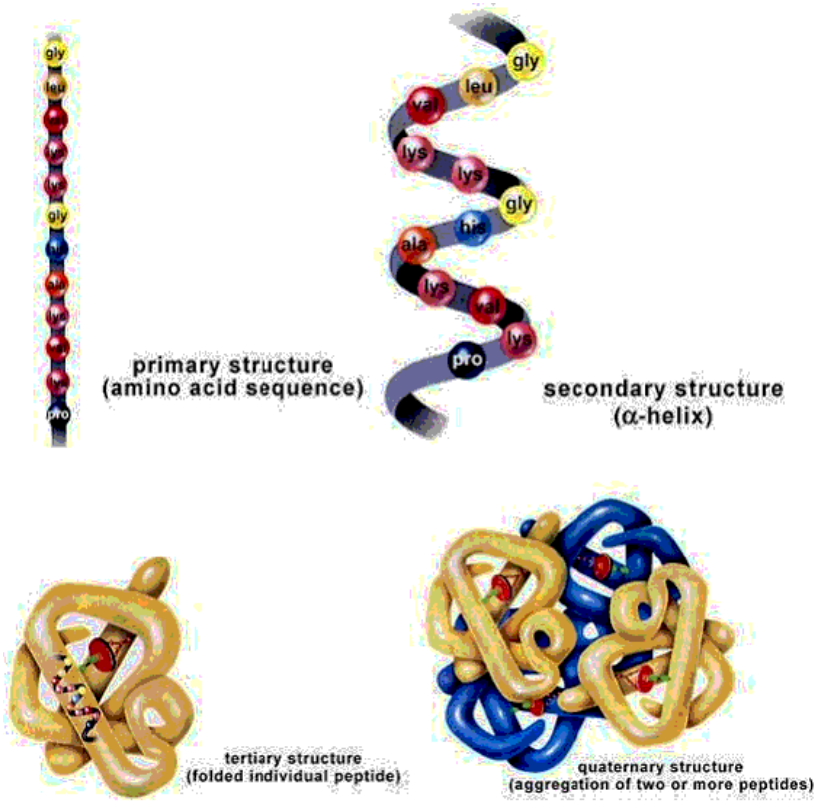

Fig. 3. The four structuring levels of the protein

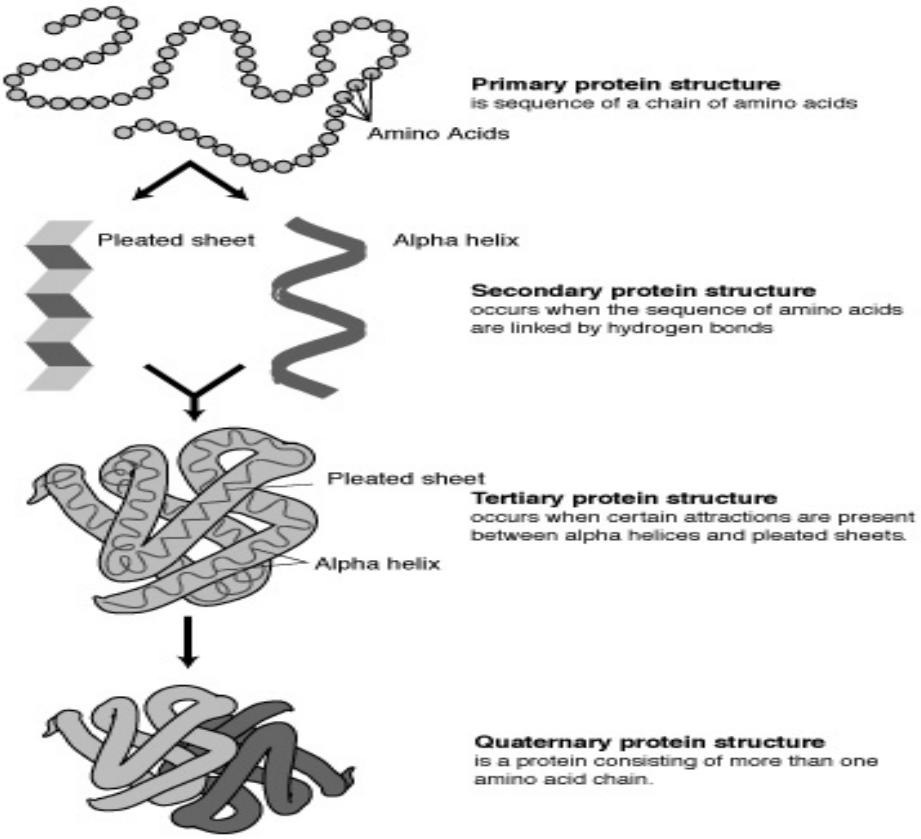

Fig. 4. Protein structure

It is important to mention that Biology and Computer Science use different nomenclature.

Here is the table that compares the notations:

Biology	Computer Science
Sequence	String, word
Subsequence	Substring (contiguous)
N/a	Subsequence
N/a	Exact matching
Alignment	Inexact matching

Subsequence (in computer science) is a non contiguous segment of a sequence. We will use the biological nomenclature. In particular, a “subsequence” will mean a *contiguous* sequence of letters.

3. Alignment

An alignment of two sequences *S* and *T* is obtained by first inserting chosen spaces, either into, at the ends of or before *S* and *T*, and then placing the two resulting sequences one above the other so that every character or space in either sequence opposite a unique character or a unique space in the other sequence. In the alignment model each two character alignment and character - space alignment is given a score (weight). Usually, insert and delete (indel) operations (alignment of a character and a space) are given the same score. Using alignment algorithms, we search for the minimal scoring (or the maximum negative scoring), representing the minimal difference or maximum similarity between the two sequences. Biological models consider the significance of each mutation and score the alignment operations accordingly. Therefore, the alignment distance can be used to estimate the "biological difference" of two DNA or protein sequences. The substitution matrix $S(i, j)$ represents the weight of each possible alignment [4].

Example The aligned sequences:

```
SEQ 1  GTAGTACAGCT-CAGTTGGGATCACAGGCTTCT
       ||||  ||  |||  ||||||    ||||||  |||
SEQ 2  GTAGAACGGCTTCAGTTG---TCACAGCGTTC-
```

Distance 1 - match 0, substitution 1, indel 2 $\Rightarrow$ distance = 14.

Distance 2 - match 0, d(A,T)=d(G,C)=1, d(A,G)=1.5 indel 2 $\Rightarrow$ distance = 14.5.

Similarity - match 1, substitution 0, indel -1.5 $\Rightarrow$ similarity = 16.5.

General setup - substitution matrix S(i,j), indel S(i,-) or S(-,j).

4. Problem formulation

The protein folding problem is reduced to transform information entities. The input is a string of characters drawn from an alphabet of 20 letters. In the simplest case, the desired output annotates each character with three numbers, giving its XYZ coordinates in the protein's three-dimensional folded shape. These coordinates are unique and depend only on the input string. There, protein structure prediction from sequence simply transforms implicit information into an explicit final form.

It is well known that the protein folding problem is also the premiere computational problem confronting molecular biology today and it is the second half of the genetic code. It is important because the biological function of proteins (enzymes) underlies all life, their function is determined by their three-dimensional shape, and their shape is determined by their three-dimensional shape, and their shape is determined by their one-dimensional sequence. The importance of computational solution is escalating rapidly due to explosion of sequences and genomes becoming available, compared to the slow growth in the number of experimentally determined three-dimensional protein structures.

The problem is unusually accessible to computer scientist because it is a pure information processing transformation, from implicit to explicit. No single computer

program would so transform the face of experimental molecular biology practice today as one that correctly, reliably, and rapidly computed this function. This is a grand challenge problem for computer science.

The problem although simply stated, is quite difficult. The process by which nature folds the string is complicated, poorly understood, and most likely the global sum of a large number of weak, local, interacting effects. Quantum mechanics provides a solution in principle, but the computation becomes intractable when confronted with the many thousands of atoms comprising a protein.

The direct approach to protein folding, based on modeled atomic force fields and approximations from classical mechanics, seeks to find the folded conformation having minimum free energy. This is difficult because a folded protein results from the delicate energetic balance of powerful atomic forces and because the vast number of possible conformation poses a formidable computational barrier.

The forces involved are often difficult to model accurately, and include stabilizing and destabilizing terms making large contributions of opposite sign summed over a very large number of atoms. Thus, small cumulative approximation errors may dominate the smaller net stabilization. For technical reasons it is difficult to model surrounding water properly, yet hydrophobic collapse is believed to be the main effect driving protein folding. Classical macroscopic parameters such as the dielectric constant become problematic at the atomic level. We may not know the protein's cellular folding context, which may include chaperone proteins, post-translational modifications, and hydrophobic interfaces to which the protein conforms. The search space [6] may exceed 10^{50} plausible folded conformations even for medium-size proteins. Simulation time-steps are measured in femtoseconds while folding time scales of interest are measured in milliseconds, a ratio of 10^{12}. Unless sophisticated methods are used, the basic time-step computation is $O(N^2)$, where N may approach 10^6 atoms with surrounding water. The simulation time may exceed 10^{12} CPU-years at current supercomputer speeds. The direct approach has been applied successfully to smaller molecules, but as yet faces stiff challenges for large proteins, though recent versions using cruder force fields are promising.

One important alternative approach is to use the wealth of information contained in already known protein structures. The structures can serve as spatial folding templates, impose constraints on possible folds, and provide geometrical and chemical information. This is an attractive strategy because proteins exhibit recurring patterns of organization; there are estimated to be only around 1 000 to 10 000 different protein structural families.

In this approach, the known structure establishes a set of possible amino acid positions in the three dimensional shape. These template spatial positions generally include only the backbone atoms, though sometimes the implied beta carbon is used as well. The highly variable surface loops are not included in the template positions. Based on topological and physicochemical criteria, an alignment of an amino acid sequence to the set of positions in one such core template is chosen. Each amino acid of the sequence is given the three dimensional coordinates of the template position to which it is aligned. Estimation of the complete structure still requires some means of assigning positions to the amino acids in the loop regions, of assigning amino acid side chain orientations and packing, and of searching the immediate structural neighborhood for a free energy minimum.

Initially, such methods employed primary sequence string similarity between the candidate sequence and the structure's native sequence in order to perform the alignment (homology modeling or homological extension). Computing the sequence similarity yields a direct alignment of amino acids in the candidate's and structure's sequences.

In cases where the sequence similarity is high this is still the most successful protein structure prediction method known. Unfortunately, it is of limited generality because novel sequences rarely have sufficiently high primary sequence similarity to another whose structure is known. Indeed, of the genomic sequences known at present, about 40 percent have no similarity to any sequence of known function, let alone known structure.

5. Conclusion

As a conclusion, although using algorithms, mathematical models and cluster's technologies, because of the complexity of protein folding, we maight state that there is no final solution and it remains a grand challenge problem for computer science.

References

1. **http://www.elmhurst.edu/~chm/vchembook/560aminoacids.html**
2. **http://www.elmhurst.edu/%7Echm/vchembook/564peptide.html**
3. **http://gened.emc.maricopa.edu/bio/bio181/BIOBK/BioBookCHEM2.html**
4. R o n S h a m i r. Algorithms for Molecular Biology. Lecture 2: November 1, 2001 Fall Semester, Tel Aviv University, 2001.
5. **http://www.rcsb.org/pdb/**
6. **http://www.ics.uci.edu/~rickl/rickl-publications**

Структурное распределение протеинов – компьютерный проблем в молекулярной биологии

Никола Чакъров

Факултет математики и информатики, Университет "Св. Климент Охридски", 1164 София

E-mails: nikola.chakarov@gmail.com chakarov@fmi.uni-sofia.bg

(Р е з ю м е)

Проблемы, связаны со структурным распределением протеинов, являются исключительно важными, потому что биологическая активность протеинов определяется их триизмерной формой в пространстве, а она со своей стороны зависит от линейного распределения и от других факторов.

Поэтому проблема, связана с изгибанием протеинов, можно свести до преобразования информации и сравнения экспериментальной структуры со знакомой структурой в базе данных.