

On the Designing of Proteins That Target Specific DNA Sequences

Computing Beyond Silicon Summer School - Caltech, July 2002
- project report -

Mirela Andronescu¹, Tiberiu Dan Onuta², and Yingley Zhao¹

¹ Department of Computer Science
University of British Columbia
{andrones,szhao}@cs.ubc.ca

² Department of Physics
University of Notre Dame
tonuta@nd.edu

Abstract. In this paper we want to find out what are the mechanisms that we could use in order to genetically engineer proteins such that they change a gene expression. It is already known how to use genetic engineering to produce proteins with a given amino acid sequence. Once we have the appropriate design of proteins, we might be able to use them in therapies for diseases based on genes' activity.

1 Introduction

The understanding of protein-DNA interaction is crucial for prediction of DNA-binding specificity of the factors which control transcription, recombination, restriction and replication. Gene expression in organisms is controlled by a variety of regulatory proteins. Structural analysis of protein-DNA complexes has revealed that the same amino acids often interact with different bases and vice versa [1]. Thus, no general rules exist yet to explain how proteins discriminate among DNA sequences or to predict their target sites. On the other hand, the amount of structural information on protein-DNA recognition has been increasing rapidly.

DNA-protein interactions occur at specific sites of the DNA, expressing or not a gene. A gene is a sequence of DNA that codes for a diffusible product. This product may be one or several proteins, as in the case of the majority of genes, or may be a type of RNA, as in the case of genes that code for tRNA and rRNA. The main characteristic is that the product obtained after transcription diffuses away from its site of synthesis to act elsewhere [2]. The first step in gene expression is always the same: the sequence along one of its strands is transcribed into a linear molecule of messenger RNA (mRNA). For our purposes, we define a gene as being *ON* if it is *expressed*, i.e. it creates protein(s). As opposed to a gene that is *ON*, we say that a gene is *OFF* if it does not produce any protein, thus having no role in the cell.

Broadly speaking, eucaryote gene expression can be achieved at any step on the pathway leading from DNA to protein [3]. The controlling of the gene expression can be done at any of these stages as follows:

1. controlling when and how often a given gene is transcribed;
2. controlling how the primary RNA transcript is spliced or processed;
3. selecting which mRNAs in the nucleus will be exported to the cytoplasm;
4. controlling which mRNAs in the cytoplasm are translated by ribosomes;
5. degrading certain mRNA molecules in the cytoplasm;
6. selectively activating or inactivating protein molecules after they have been created.

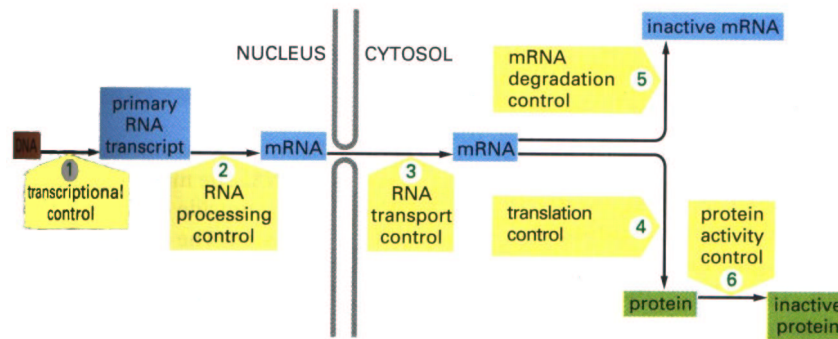


Fig. 1. The six steps at which eucaryote gene expression can be controlled.

In the process of its transcription to mRNA, the gene is helped by an enzyme called RNA polymerase. The process begins with the binding of this enzyme near the beginning of a gene to a site called *promoter*, whose length is roughly 60 bp. Each promoter points RNA polymerase in either one or the other direction along the DNA helix. While the enzyme moves to a given direction, it copies one of the strands into a new mRNA strand. The polarity of the mRNA chain is opposite to that of the DNA template strand.

RNA polymerase can be helped or hindered in its attempt to transcribe a gene by regulatory proteins that bind to sites on the DNA called *operators*. A *negative regulatory protein* prevents transcription and a *positive regulator* stimulates the transcription of a gene. A particular regulatory protein binds to a specific operator site(s) on a DNA molecule.

In this work, we want to change the role of a gene, by turning it *OFF* when it was *ON* by default, and by turning it *ON* when it was *OFF*. The **motivation** behind this work is mainly related to therapies for diseases that rely on genes' activity. We concentrate on achieving this behaviour by acting at the level of transcription initiation, the first step out of the different aforementioned stages. The most important class of transcription factors is known as zinc finger DNA-binding proteins (ZFP), that bind to the operator of a specific gene. The ability to engineer zinc fingers to specifically turn *ON* or *OFF* a gene function was first discovered by Carl Pabo at the Massachusetts Institute of Technology. Sangamo

BioSciences Inc. has also created ZFP transcription factors that can control gene expression, and consequently, cell function [5].

The remainder of this paper is organized as follows: section 2 introduces three related works that are relevant for our purposes. In section 3 we try to understand the factors that are definitory for protein-DNA interactions, focusing on zinc fingers and publicly available protein databases. We also propose some novel ideas on how to design DNA-binding proteins to turn a specific gene *ON* or *OFF*. Finally, section 4 presents conclusions and gives ideas of future work.

2 Related Work

Sarisky et al. [6]

Sarisky et al. [6] created a tool called *ORBIT* (Optimization of Rotamers by Iterative Techniques) to design proteins that target specific DNA sequence. Using the yeast transcription factor GCN4 as a model system, they first computationally generate sequences predicted to bind with high affinity to the wild type DNA target. The proteins are expressed in *E. coli*. and experimentally characterized by gel shift electrophoresis and DNaseI foot printing.

They elucidated some of the factors which are important for DNA binding with site-specific recognition as well as to develop a methodology for generating novel proteins with high affinity for target DNA sites. They developed a general method to computationally select protein sequences that recognize a target DNA sequence. They use a force field developed for use in the protein design process ORBIT. The force field used in ORBIT includes terms for van der Waals contacts, solvation (a benefit for burial of hydrophobic surface area, a penalty for burial of polar surface area or exposure of hydrophobic surface area), electrostatics, and hydrogen bonding.

They used a force field and a design methodology that allow generation of protein sequences to bind to any target DNA sequence.

They describe a general methodology of generation of novel proteins that bind site-specifically to DNA. This approach will allow us to elucidate the relative importance of various force to the formation of protein-DNA complexes. The ability to target specific DNA sequence with small proteins may also prove useful for therapeutic or diagnostic purpose where binding of a specific DNA sequence is necessary.

Jamieson et al [7]

In [7], Jamieson et al. use random mutagenesis and phage display to alter the DNA-binding specificity of Zif268, a transcription factor that contains three zinc finger domains. Four residues in the helix of finger 1 of Zif268 that potentially mediate DNA binding were identified from an X-ray structure of the Zif268-DNA complex. A library was constructed in which these residues were randomly mutated and the Zif268 variants were fused to a truncated version of the gene III coat protein on the surface of M13 filamentous phage particles.

They were unable to enrich for clones that bind to five other binding sites (ACG, CCG, CGC, ATA and TAT), suggesting modification of just these four

residues in finger 1 may not allow it to adapt to all DNA binding sites. The studies show that it is possible to isolate zinc fingers by Phage display that distinguishes operator sequences that differ by a single base change. Moreover, such selection methods should aid in clarifying rules for zinc finger-DNA recognition.

The paper suggests that phase display is a powerful tool for sorting libraries of zinc finger proteins for binding to different operator sequences. Specific enrichments were seen after several rounds of sorting with some operator sequences. When clones were analyzed after several rounds of selection for a give operator, they often had similar sequences and differed from clones isolated using different operator sequence. Moreover, some operator sequences gave no enrichments for binding phage, and when sequences from one of these (ACG) were analyzed, little consensus was observed.

Although the specificity advantages determined for the few clones analyzed are mild, the clones were still powerful selected. There are limitations of the phage display method that need to be emphasized.

Bulyk et al. [8]

In [8], Bulyk et al. describe a DNA microarray-based method to characterize sequence-specific DNA recognition by zinc-finger proteins. A phage display library, prepared by randomizing critical amino acid residues in the second of three fingers of the mouse Zif268 domain, provided a rich source of zinc-finger protein with variant DNA-binding specificities. Microarrays contain all possible 3-bp binding sites for the variable zinc fingers permitted the quantitation of the binding site preferences of the entire library, pools of zinc fingers corresponding to different rounds of selection from this library, as well as individual Zif268 variants that were isolated from the library by using specific DNA sequence. The results demonstrate the feasibility of using DNA microarrays for genome-wide identification of putative transcription factor-binding sites.

They use DNA microarray to examine the spectrum of binding-site specificities of a collection of Zif268 mutants selected from a phage display library of the second finger. Quantitative measurements of more than 750 DNS-protein interactions were gathered from 10 different microarray-binding assays by using wild-type Zif268, four mutants and seven pools of mutants.

3 Description

As we mentioned in the Introduction, a *negative* control of the gene is possible, as well as a *positive* one [2]:

- A *negative control* prevents a gene from being expressed. Figure 2 shows a gene that has the state *ON* by default. The transcription is initiated by RNA polymerase that recognizes the gene's promoter. If a *negative regulatory protein*, also called *repressor*, perfectly binds to the operator, RNA polymerase cannot initiate the transcription, therefore turning the gene expression *OFF*.
- RNA polymerase sometimes cannot initiate the transcription by itself. In this case, a *transcription factor* is required to assist it in binding to the promoter, having an important role in the *positive control* of the gene. The gene showed

in Figure 3 is inactive by default. Several factors are needed to bind to sites in the vicinity of the promoter, in order to enable RNA polymerase to initiate the transcription. The ultimate effect of these factors is thus to turn the gene *ON*.

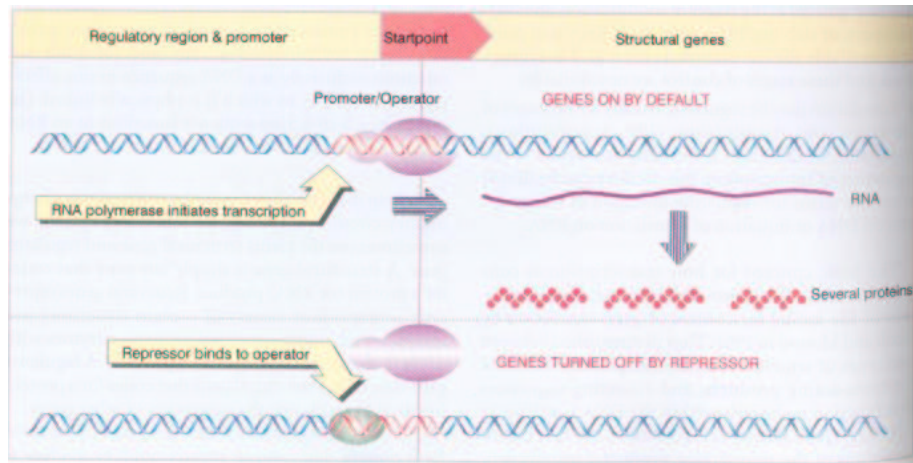


Fig. 2. In negative control, a repressor binding to the operator of the gene can turn the gene *OFF*.

Proteins fold into characteristic shapes, determined by their amino acid sequences. Different amino acid sequences usually fold into different shapes, and protein folding is a hard and yet unsolved problem. Nevertheless, small structural motifs are found in many proteins, the most common and studied one being the α helix. Figure 4 shows an α helix that fits perfectly into the major groove of a DNA strand.

3.1 Regulatory Proteins

The transcription factors have been researched and analysed. Therefore, we can compare them and find that common types of *motifs* are responsible for binding to a DNA molecule. The motifs are short, comprised a small part of the protein structure, and are also responsible for activating transcription via interactions between proteins of the transcription machinery. There exists detailed information about several groups of proteins that control transcription by using different motifs to bind to DNA molecule.

A classification of DNA-binding protein structures includes:

1. *Zinc containing DNA binding domains*
 - *Zinc fingers* are comprised of an α helix and a β strand. Zn binds two histidines in the α helix and two cysteines in the β strand. Basic amino

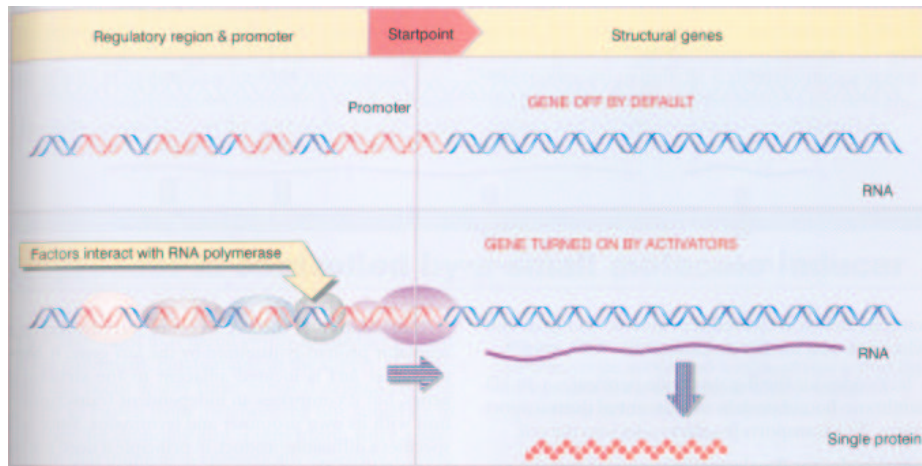


Fig. 3. In positive control, certain factors are needed to help RNA polymerase to initiate the transcription, thus turning the gene *ON*.

acids in the α helix contact nucleotides in the major groove. Each of the three fingers (see Figure 5) from Zif268 contact nucleotides and phosphates via basic residues. The β strand of Zn fingers positions the α helix in the major groove. Zn finger proteins with multiple fingers typically bind as monomers.

- *Ga14* binds a upstream activating sequences as a dimer. It uses six cysteines and two Zn ions to bind DNA. Basic residues in the α helix makes contacts with bases in the major groove. Dimerization motif is comprised of a coiled α helix.
- *Nuclear receptors* interact with hormone ligands and bind to hormone response elements to activate transcription. The nuclear receptors are always in the nucleus and bind target sequence and repress transcription

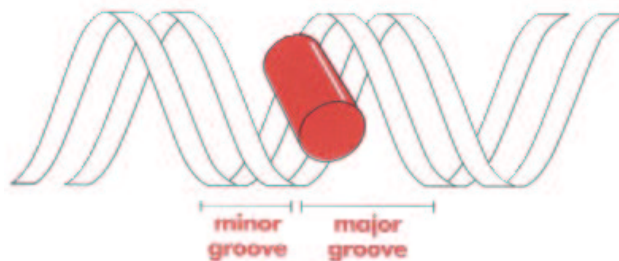


Fig. 4. Minor and major grooves of a DNA strand, and an α helix site of a protein, that fits the major groove.

in the absence of hormone, but activates transcription in the presence of hormone. Binding occurs at half sites separated by 3bp using the recognition α helix.

2. *Homeodomains* are comprised of three helices where the first two form a *helix-turn-helix* and the third is the recognition helix. They have weak binding specificity on their own.
3. *bZIP* and *bHLH* domains both domains combine DNA binding and dimerization. The "b" in the front of acronyms stands for the basic DNA binding. The ZIP domain, also called *leucine zippers*, is an α helix having leucines every seven amino acids, so they lie on the same side of the helix. Dimerization is mediated by a "coiled coil" formed by interactions among leucines. bZIP binds DNA like forceps. bHLH (helix-loop-helix) also forms dimers via α helix interactions which position the basic region in the major groove. There is an independence of domains, that is, the activation and DNA binding domains are modular and can be swapped to change the specificity of binding and activation or repression.

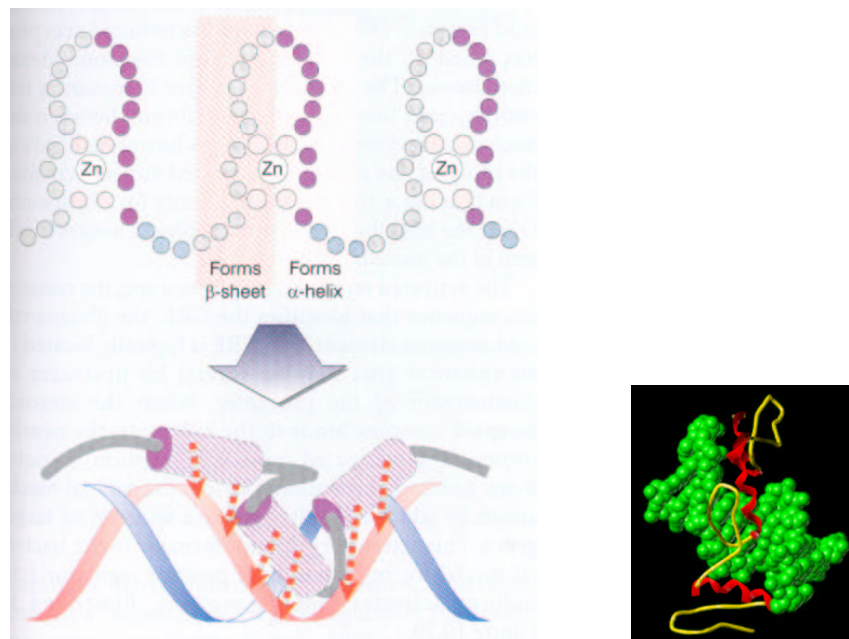


Fig. 5. Zif268 Zinc Finger-DNA complex (1AAY): a schematic view and a 3D figure

3.2 Physical Bonds between proteins and DNA

The binding between a DNA strand and a protein(s) is on hydrogen-bond bases (see Figure 6). Direct evidence for the existence of individual hydrogen bonds

in such biomolecules has been provided by the detection of hydrogen bond scalar couplings in DNA, proteins, and their complexes. These scalar couplings are electron-mediated and can be used to identify the three partners involved in the hydrogen bond, i.e. the donor, the acceptor and the proton. The size of the hydrogen bond scalar couplings correlates with the strength of the hydrogen bond and with the chemical shift of the proton involved in the H-bond. Because of the characteristics of the hydrogen bonds, the proteins have high affinity for a specific DNA sequence and a low affinity for any other DNA sequence. Actually, there is only one high-affinity site in the gene: the operator. The remainder of the gene provides low-affinity binding sites.

In the DNA-protein interactions, the hydrogen bonds appear between hydrogen atoms belonging to DNA bases (or amino acids from protein) and N or O atoms from amino acids (or DNA bases).

In addition, there are other standard force fields that should be taken into account in order to explain the specificity of DNA-protein interactions. In some models, it turns out that favorable van der Waals terms have the role of overpowering the hydrogen bonding terms. However, this is unlikely to be true, since the specific hydrogen bonds are more valuable than non-specific van der Waals contacts for specific binding.

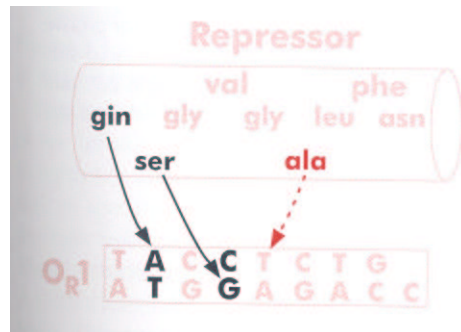


Fig. 6. An example of physical bonds between a protein and a DNA sequence

3.3 Examples of proteins that recognize specific DNA sequences

Genetic analyses have started in bacteria since 1950s, providing the first evidence of the existence of gene regulatory proteins, many such proteins being discovered. Hundreds of DNA sequences have been also identified, each recognized by a different gene regulatory protein or by a set of related gene regulatory proteins. Table 1 shows some examples of such proteins, the length of their amino acid sequence, their code in the Protein Database [9], along with the DNA sequences that they recognize [3].

Organism	Protein name	No. AA	PDB Id [9]	DNA sequence recognized
Bacteria	lac repressor	170	1L1M	AATTGTGAGCGGATAACAATT TTAACACTCGCCTATTGTAA
Bacteria	lambda repressor	87	1lmb	TATCACCGCCAGAGGTA ATAGTGGCGGTCTCCAT
Yeast	GAL4	170	1D66	CGGAGGACTGTCCTCCG GCCTCCTGACAGGAGGC
Mammals	GATA-1	76	1GAT	TGATAG ACTATC

Table 1. Some Gene Regulatory Proteins and the DNA sequences they recognize

3.4 Data Banks

Several databases that show protein-DNA interactions currently exists, being updated continuously.

Protein Data Bank [9]

The Protein Data Bank was established at Brookhaven National Laboratory [10] in 1971 as an archive for biological macromolecular crystal structures. Since October 1998, the PDB has been managed by the three members of the Research Collaboratory for Structural Bioinformatics (RCSB) - Rutgers, The State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego, and the National Institute of Standards and Technology.

PDB contains techniques of X-ray crystal structure determination, NMR, cryo-electron microscopy and theoretical modeling.

Nucleic Acid Database [11]

The Nucleic Acid Database (NDB) was established in 1991 as a resource to assemble and distribute structural information about nucleic acids. Over the years, the NDB has developed generalized software for processing, archiving, querying and distributing structural data for nucleic acid-containing structures.

Structures available in the NDB include RNA and DNA oligonucleotides with two or more bases either alone or complexed with ligands, natural nucleic acids such as tRNA and protein-nucleic acid complexes. The archive stores both primary (such as crystallization information, data-collection etc.) and derived information about the structures.

Protein-Nucleic Acid Complex Database [12]

This database contains structural data of protein-nucleic acid complex, classified according to recognition motif of proteins and DNA forms involved in the complex. It also helps researchers to understand the relationship between the structure, property and function of biomolecules. The structural data are taken from Protein Data Bank [12], and implemented into a relational database.

Valuable research has been done to understand the protein-nucleic acid interactions. Researchers collected experimentally observed thermodynamic data of protein-nucleic acid binding, and created a relational database called ProNIT [13], that is available on-line. It contains several important thermodynamic data for protein-nucleic acid binding, as well as structural and quantitative binding data.

The database includes protein-nucleic acid binding data from scientific articles, and it currently contains 2514 entries. Using ProNIT, it would be possible to establish a relationship between thermodynamics and structural changes upon the formation of protein-DNA complexes [13].

4 Conclusions and Future Work

In this work, we tried to find the mechanisms based on DNA-binding proteins that could change a gene expression, and also the specific factors that influence this machinery. Future work would imply a better understanding of the protein-DNA interactions and eventually the creation of a computer program able to design one or more proteins that would change the expression of a given gene.

Hydrogen bonds implied in protein-DNA interactions are characterized by a scalar force field. The latter can be characterized by a potential function, whose form is roughly known. As this potential field is, in our goal, an essential part for computing, we will try different mathematical forms such as: Toda potential, modified Toda potential, Morse potential and Lennard-Jones potential. As soon as we obtain some data using these theoretical models, we will compare them with the experimental results from the public databases (such as ProNIT etc.). Thus, we want to figure out which one of the aforementioned potential forms is the most appropriate for our goals. As soon as we know this, we can create a computer program to design proteins that target to the given DNA.

References

1. L. A. Mirny, M. S. Gelford, *Structural Analysis of Conserved Base Pairs in Protein-DNA Complexes*, Nucleic Acid Research, 2002, Vol. 30, No. 7, p. 1704-1711.
2. B. Lewin, *Genes VII*, Oxford University Press, 2000, p. 231-318, 649-684.
3. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. D. Watson, *Molecular Biology of the Cell*, Third Edition, p. 401-417.
4. Mark Ptashne, *A Genetic Switch*, Second Edition, p. 33-48.
5. Sangamo BioSciences Inc. <http://www.sangamo.com>
6. Sarisky et al., *Computational design and experimental validation of novel DNA binding*.
7. Jamieson et al., *In Vitro Selection of Zinc Fingers with Altered DNA-Binding Specificity*, Biochemistry 1994.
8. Bulyk et al., *Exploring the DNA-binding specificities of zinc fingers with DNA microarrays*.
9. Berman et al., *Protein Data Bank*, Acta Cryst. (2002). D58, 899-907
<http://www.rcsb.org>
10. Brookhaven National Laboratory <http://www.bnl.gov/>
11. H. M. Berman, J. Westbrook, Z. Feng, L. Iype, B. Schneider and C. Zardecki, *The Nucleic Acid Database*, Acta Cryst. (2002). D58, 889-898

12. Protein-DNA Recognition Database
<http://www.rtc.riken.go.jp/jouhou/3dinsight/recognition.html>
13. Prabakaran et al., *Thermodynamic Database for Protein-Nucleic Acid Interactions*,
Bioinformatics, Vol. 17 no. 11 2001, p. 1027-1034.
<http://www.rtc.riken.go.jp/jouhou/pronit/pronit.html>
14. Protein Information Resource <http://pir.georgetown.edu>