The Atom of Evolution

Jonghwa Bhak[1,2,3,4 *], Dan Bolser[3,5], Daeui Park[4], Yoobok Cho[4], Kiesuk Yoon[4], Semin Lee[1], SungSam Gong[1], Insoo Jang[2], Changbum Park[1], Maryana Huston[3] and Hwanho Choi[3]

[1]Biomatics Lab, BioSystems, KAIST, Daejeon 305-701, Korea, [2]NGIC, KRIBB, Daejeon 305-333, Korea, [3]BiO Institute, Daejeon 305-701, Korea, [4]OITEK Inc. Daejeon 305-701, Korea, [5]MRC-DUNN, Cambridge CB2-2XY, England, United Kingdom.

Abstract

The main mechanism of evolution is that biological *entities* change, are selected, and reproduce. We propose a different concept in terms of the main agent or atom of evolution: in the biological world, not an individual object, but its interactive *network* is the fundamental unit of evolution. The interaction network is composed of interaction pairs of information objects that have order information. This indicates a paradigm shift from 3D biological objects to an abstract network of information entities as the primary agent of evolution. It forces us to change our views about how organisms evolve and therefore the methods we use to analyze evolution.

Introduction

Bioinformatics and paradigm shift.

Bioinformatics is a scientific discipline that analyzes, seeks understanding of, and models all life as an information processing phenomenon on utilizing energy with methods from philosophy, mathematics, and computer science using biological experimental data. Due to its information processing nature, bioinformatics is one of the broadest and deepest scientific disciplines. All biological research is, aiming to understand the architecture of information processing in life. For example, the analysis of genomes and genes is to

*Corresponding author: E-mail j@bio.cc,
  Tel +82-42-869-4318, Fax +82-42-869-4310
  Accepted 3 October 2004

discover the underlying linguistic rules of molecules (Searls, 1993) and the concept of proteins as computational elements (Bray, 1995). In this sense, the shortest possible definition of bioinformatics is: Bioinformatics is Biology and Biology is Bioinformatics. In the next 20 years, bioinformatics will become the core of biological education even in secondary school biology courses. The early 2000s is at the point where the conventional views of molecular biology should change with new revolutionary views in biology. One of them is the transition from the conventional object-oriented understanding of biology to an *interaction-oriented* understanding.

Objects and interaction

The concept of evolution is the backbone of modern biology (Darwin, 1859). Evolution is used to explain development and change in many different areas of human society. The word evolution is from the Latin evolvere meaning "unfold". It succinctly means "descent with modification of objects", whatever the objects are. For example, although a car is not a self-replicating organic organism, an evolutionary process appears to have occurred in car models. The cars were duplicated, changed, and selected as a result of the interaction of manufacturers and users: by observing the users' selection and through market research, the manufacturers made design decisions for each new model. As abstract information objects, the cars evolved as much as human hair evolved over time. Hair is a product of hair cells while cars are the product of human brain cells. Here, the more influential factors in development are the interaction and information rather than the objects themselves. This leads us to the question of what the fundamental evolutionary unit is in interaction processes.

Discussion

1. The fundamental unit of evolution is the interaction pairs of biological information objects.

Among many biological information object levels, cell-cell interactions are not the most basic unit of evolution, as non-cellular entities, such as molecules and viruses, show evolutionary processes. Use of the

Fig. 1. Two interacting molecules: a protein complex of two structural domains. Domains are perhaps the most fundamental biological unit giving any meaning or function in life.

molecular level was our choice out of a number of possible levels found in the evolution of life. Unlike cells, genes and proteins at the molecular level contain important biological information as ordered data. The interaction process we suggest as the most basic unit of evolution lies at the molecular level where individual genes and proteins interact as special information objects with consistently ordered sequences (Fig. 1). Below and above this level, it is not easy to define structural or sequence order information.

2. Interaction networks

An interacting molecular pair with internal order information is the most basic unit of evolution with the potential of being selected. An interaction network is defined as a system that maintains its pairing architecture over time. All information objects, such as genes and proteins, "know" precisely with what other molecules they should or should not interact within that network. The interaction network as the sum of molecular interaction pairs, is the most conserved information entity in biology. We will discuss several issues and research development toward molecular interaction network in the following sections.

2.1. Misinteraction: Many diseases including BSE (bovine spongiform encephalopathy or mad cow disease) are the result of the wrong interaction of proteins. that is said to be the cause of BSE is, in fact, a kind of misinteraction. In the evolution of proteins, avoiding misinteraction is extremely important. Most venoms and toxins work by causing such misinteractions in cells. For many proteins' evolution, it is perhaps more important to avoid undesired interactions than to find the desired interaction partners.

2.2. Complexity of cells and cell compartmentalization caused by the limited interaction specificity of protein

structures: There is nearly an infinite number protein sequences. At this very moment, new proteins are being synthesized by some unknown organisms as constant mutations occur in nature. Due to this ongoing process, it is theoretically not possible to map all the protein sequences. However, there seem to be fewer than 2000 distinct protein folds that represent the infinite number of protein sequences. In a small bactrium, 2000 types of interaction nodes can have enough specificity to avoid misinteraction. However, in a large eukaryotic cell, with the given 3D molecular objects, it is difficult to avoid misinteraction. A protein domain can have several different specificity and specific interaction interfaces; however, with many paralogs within a single-cell bag, it still reaches the theoretical limit to control misinteractions. Inevitably, this selection force, in order to have more diverse interaction types and their functions, drove cells to compartmentalize. This is an object and function encapsulation to make higher complexity information architecture.

2.3. Protein folding is a molecular interaction problem: Misinteraction between amino acid residues within a protein leads to problems in the protein folding process in general. Protein folding in action is an interaction problem. Protein regions and amino acid regions interact with other regions and residues to form a network that is evolutionarily conserved due to order information kept over billions of years. The protein folding problem can be divided into local and long-distance interaction problems. The local interaction problem is the secondary structure formation and prediction problem. How amino acids interact locally determines the alpha helix, beta sheet, and other structures. The regularity is such humans can predict the secondary structures very reliably using algorithms such as neural networks. However, no algorithms have yet reached an average accuracy over 80%. This is because the long distance interaction information of proteins is missing.

2.4. Protein folding interaction type ratio is 79: 21: Therefore, one can hypothesize that the ratio between the local and long-distance molecular interaction ratio in protein folding is about 80:20. A test was carried out to measure the correct ratio using the reversed sequence set of protein structures from PDB (Protein DataBank: http://rcsb.org). It turned out that 79% of the protein structures have the same secondary structures (local residue interaction) regardless of the order of the protein sequences (Park et al., 2000). Therefore, the precise ratio of the local and long-distance molecular interaction ratio in protein folding is suggested to be 79:21. By
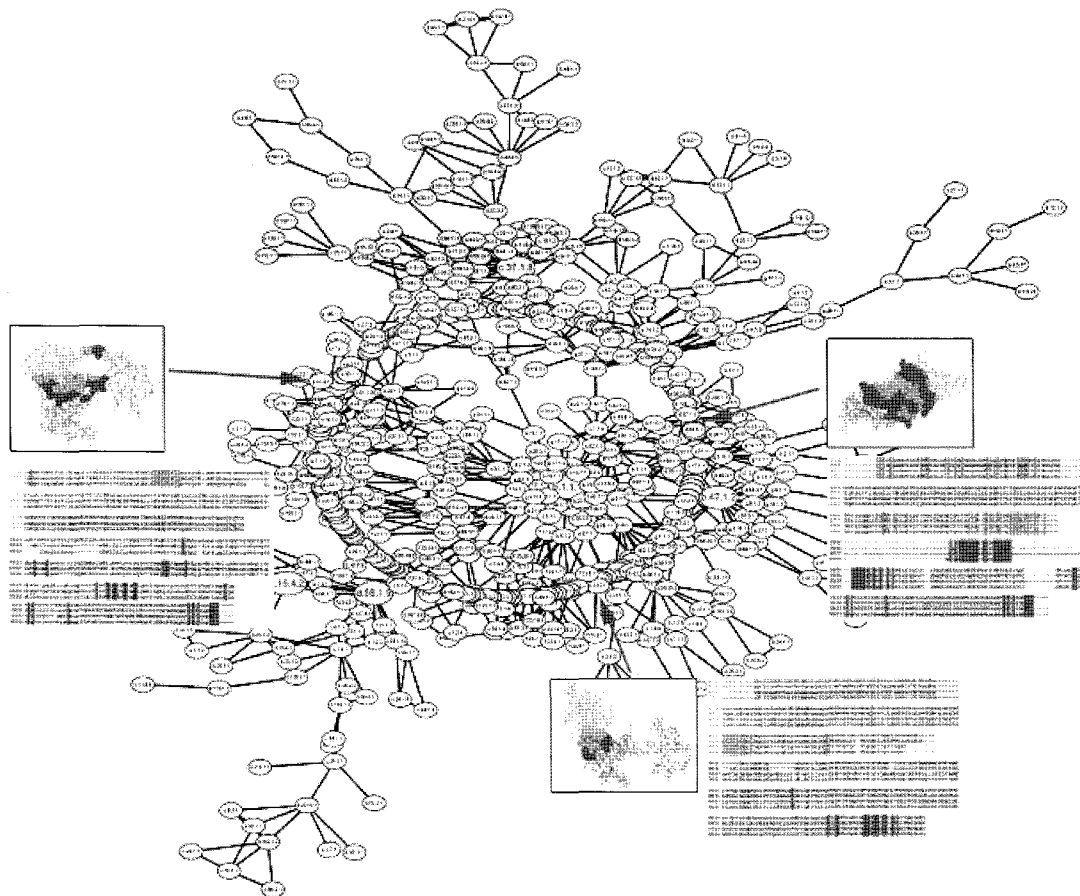
Fig. 2. An interaction network. Nodes are interacting protein domain families. Nodes with red characters are hub nodes. The multiple alignments show the interaction interfaces that belong to the interfaceome of proteins. The 3D pictures show the interacting domains with their interfaces highlighted.

devising a method to detect 21% of the long-distance interaction network information, one may be able to solve the protein folding problem computationally.

2.5. Interaction network is tight: Due to the risk of misinteraction, biological networks are extremely conservative (Park and Bolser, 2001). The interaction backbone of any interactome is very similar to all other known life forms. With the analysis of over 140 genomes, it is clear that there is a unique interactome core found by a comparative interactomics (manuscript in submission). The core interaction has the most critical functions of life such as RNA binding, energy metabolism, and translation machineries.

2.6. Driving of evolution by old and important families: The tightness of the interactome can also be demonstrated by the small number of protein families that seem to be the major driving force of species

evolution. Their most representative is the PLCH (P-loop containing hydrolase: c.37.1 in SCOP) domain family, found at numerous important protein interactions in the widest range of species known (hence, perhaps the oldest and most important protein fold). The second most prominent family is the immunoglobin family that has prospered since the appearance of eukaryotes. While PLCH domains are found in almost all important functions, including energy metabolism and many enzymatic functions, immunoglobin famlies occur more in structural functions. A different niche or territory was occupied by the two different structures.

2.7. The age of protein folds and biological species measured by comparative interactomics: Measuring the age of any particular species is more a philosophical than a biological problem. All genomes have a complication in their genome origin due to horizontally transferred genes, deletions, and insertions that mask

their lineage. Calculating the precise age of any species is theoretically impossible. Practically, there are no fossils that can reveal the exact age of any microbial organisms that lived billions years ago (Schof, 1993). However, it is possible to measure the relative antiquity of species by measuring the relative age of protein families. In the past, protein sequence alignments were used to estimate the rough date of the divergence of species (Doolittle et al., 1996). However, using an individual protein's divergence is not sufficiently precise. Using protein structural domains and families is a more reliable method. Protein domains and families have a proportionally linear relationship with the evolution of cells and species. Their rate of evolution is strikingly constant. However, the repertoire of protein families alone cannot explain the diversity of life forms. One can compare small RNA molecules to derive evolutionary distances among species. However, organisms have systems level differences in their morphology (for example, large multicellular organism versus microscopic monocellular organisms) and mode of life. Therefore, the best approach presently available is to compare their interactomes. Interactomes can reflect the complexity of the organisms. To answer a simple question of how one species differs from another, the difference of interaction networks of all the molecules can be an answer. Measuring the age of protein families can be done by the occurrences of proteins in presently known organisms. Proteins that are found in many species will have a greater chance of being older than others (Bolser and Park, 2003).

2.8. Evolution of interaction networks: Analyzing the pairs and networks in evolution is not easy, as there is a limited amount of molecular interaction information at present. Also, there are multiple layers of interaction. Protein domain-domain is a fairly distinct layer of molecular interaction. It is particularly useful to study the long evolution of protein structures. Protein-protein layer interaction that is one level higher than domain is important in real-time cellular processes. Furthermore, most proteins function within or in association with protein complexes that are at another layer of molecular interaction. Describing such multi-scale molecular interaction is not straightforward (Moon et al., 2005). An attempt has been made to map protein domain-domain interactions on a global scale from a 3D structural database (Park et al., 2001) which could map the interaction of structural groups or protein families. One can analyze the evolution of interaction networks using the bioinformatic homology assignment of various genomes (Lappe et al., 2001; Bolser et al., 2003; Kim et

al., 2004). Computationally, by using the fused gene, it information, it is possible to detect the domain-domain I interaction (Marcotte et al., 1999; Enright et al., 1999). On the experimental side, large scale yeast two hybrid analysis enabled interactomes of several genomes. These approaches use full protein-protein interaction (Ito et al., 2001; Uetz et al., 2000). The advantage of using these approaches is that the information reflects each species' molecular interaction as a complex network, as long as the experimental technique (Raicu et al., 2004) is reliable. Yeast two hybrid experiment resembles a snapshot of the interactomes along evolutionary time. Recently, more research results have been accumulated on protein domain (fold) structure interactions (Chia and Kolatkar, 2004)

2.9. Interactomes and interaction maps: The term interactome has been used independently by several bioinformatics researchers from the late 1990's (http://interactome.org). Around the same time, on the experimental side, very large scale yeast two hybrid system data became available for yeast and C. elegans (Walhout et al., 2000). Now, the list includes the fly interactome (Sanchez et al., 1999). Recently, there have been comparative studies of these large scale interactomes. Using the available experimental data, other groups started to predict the interactions of proteins using probabilistic models and interaction matrices. Although there is the important problem of obtaining experimental verification for such predictive methods, they can dramatically extend the scope of data in the interactome field.

2.10. Many interactomes: Now the field is moving into a new era where people are constructing the whole human interactome (http://hpid.org; Kim et al., 2003; Lehner and Fraser, 2004). The species specific interactome concept is not new. Other interactomes, such as the mitochondrial, have been built; it was estimated that there are around 1500 protein domains in mitochondria. The rice genome (Oryza sativa) has shown 1441 protein family-family interactions from the 32% assignment rate of proteins to all the predicted and known gene sequences (http://oryzasativa.org). Lately, the scope of interactomes has gone toward disease specific proteome such as the down syndrome interactome (Gardiner et al., 2004).

2.11. Interfaceome: the puzzle pieces of interaction: As one can describe protein folding as the interaction network of small molecules such as amino acids, the protein interactome can be subdivided into many pieces of interfaces. Mapping all such small pieces leads to

structure based computational drug discovery. Interfaces are the bolts and nuts of the evolution, if interaction is the atom of evolution as a concept and principle. Interfaces do not contain any order information, therefore, unlike protein sequences, it requires a different representation method. A traditional way to do this is to represent them as 3D coordinates that can be grouped and belonged to a protein domain. This enables a hierarchy from the highest possible protein structure to its interface. For example, in SCOP hierarchy, interface X can be classified as **C.371.1.1.X**. It belongs to the SCOP class C and superfamily 37, family 1, domain 1, and the interface X. The consequence is that we can map all such structural interfaces into an organized network or map called *Interfaceome* (http://interfaceome.org). There has been much research on individual and groups of interfaces (Chothia and Janin, 1975; Lawrence and Colman, 1993; Jones and Thornton, 1997; Jones and Thornton, 1997; Jones et al., 2000). Interfaceome aims to cover all the interfaces in a network with a clear hierarchy that can be represented in many different ways such as in a Voronoi diagram (Richards, 1974; Richards, 1977; Poupon, 2004). Protein interfaces can be detected by simple distance measurement (PSIMAP) and accessible surface area (ASA, Chothia, 1976, Miller et al., 1987). InterPare is a database server that provides interfaceome information (http://interpare.net).

3. Network Biology

3.1. Network evolution: The term network evolution indicates that the agent of evolution is the network rather than individual components of the network such as proteins and genes related to aging process (Promislow, 2004). It has been found that interaction networks, perhaps all biological networks, are very conservative and tight in evolution (most important biochemical pathways and regulation networks are shared by species in different superkingdoms). New molecular interactions
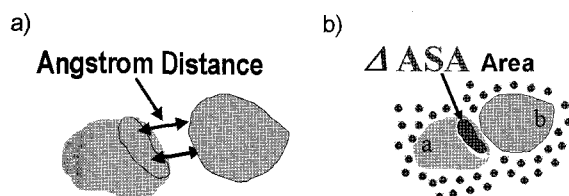
a)   b)

Fig. 3. Determining interaction. a) Simple distance measure (for example, 5 angstrom 5 contacts) among proteins or protein domains. b) Accessible surface area (ASA) method for determining the interfaces. Delta ASA is the difference of ASA between bound and unbound forms of proteins.

are carefully chosen, and perhaps the core interactome has not changed since the very beginning of life. Network biology (http://networkbiology.org) is a new discipline where biological research is carried out with networks as the central perspective (Jeong et al., 2001; Albert and Barabasi, 2002; Albert and Othmer, 2003; Ng and Tan, 2004). Essentially, all biological entities (bio-entities) are analyzed as networks with their modular components (Gagneur et al., 2004). With this new perspective, the most critical aspect of biological problems is the information process capability and information processing architecture as represented in networks. One of the outcomes of network biology is the calculability of the complexity of life. Complexity is a measure of the concentration of information, and life can only be completely analyzed by mapping the architecture of complexity.

3.2. Network visualization: One of the first steps of analyzing complexity is representing data with visual tools such as Pajek (Batagelj and Mrvar, 2001), InterViewer (Ju et al., 2003), ProViz (Iragne et al., 2004), and PINC (Hongchao et al., 2004).

Conclusion

Modern biology is becoming an information science as contemporary biological problems are associated with networks, circuits, controls, and molecular evolution. An interaction pair as the most fundamental evolution unit has been introduced for mapping interactomes and biological networks. The ultimate understanding of life will be achieved through understanding the complexity of such interaction networks at two very different domains: a long evolution interaction process and a short period of cellular interaction processes. Recent efforts on mapping the interactions of biological information objects, such as interactome map, interfaceome map, network biology approach, and various servers associated with them, are introduced. Complete computational simulation of life by human beings will be achieved in the next few decades. However complex this will be and whatever experimental techniques will be necessary for us to reach that stage of evolution, it is necessary to map and interpret the interactions of molecules, especially proteins.

Acknowlegements

References

Albert, R., Jeong, H., and Barabasi, A.L. (2000). Error and attack tolerance of complex networks. *Nature* 406, 378-382.

Albert, R. and Othmer, H.G., (2003). The topology of the regulat- ory interactions predicts the expression pattern of the seg- ment polarity genes in *Drosophila melanogaster. J.Theor. Biol.* 223, 1-18.

Albert, R. and Baraba'si, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47-97.

Batagelj, V. and Mrvar, A. (2001). Pajek-analysis and visualization of large networks. *LNCS* 2265, 477-478.

Bolser, D.M. and Park, J. (2003). Biological Network Evolution Hypothesis Applied to Protein Structural Interactome. *Genomics & Informatics* 7-19.

Bolser, D., Panos, D, Harrington, R., Park, J., and Schroeder, M., (2003). Visualisation and Graph-theoretic Analysis of a Large-scale Protein Structural Interactome., *BMC Bioinformatics* 445, 1471-2105

Bray, D. (1995). Protein molecules as computational elements in living cells. *Nature* 376, 307-312.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science* 13, 190-202.

Chia, J.M, and Kolatkar, P.R. (2004). Implications for domain fusn protein-protein interactions based on structuralio information. *BMC Bioinformatics* 26, 161 [Epub ahead of print]

Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature* 256, 705-708.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J.Mol.Biol.* 105, 1-14.

Darwin, C. (1859). On the origins of species, London, *http://pages.britishlibrary.net/charles.darwin/texts/origin 1859/origin_fm.html*

Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G., and Little, E., (1996). Determining Divergence Times of the Major Kingdoms of Living Organisms with a Protein Clock. *Science* 271, 470-477.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete

genomes based on gene fusion events. *Nature* 402, 86-90.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome, Y. *Proc. Natl. Acad. Sci. USA* 98, 4569-4574. Epub.

Jones, S. and Thornton, J.M. (1997). Analysis of Protein-protein interaction sites using surface patches. *J. Mol. Biol.* 272, 121-132.

Jones, S. and Thornton, J.M. (1997). Principle of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* 93, 13-20.

Jones, S., Marin, A., and Thornton, J.M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Engineering* 13, 77-82.

Ju, B.H., Park, B., Park J.H., and Han K. (2003). Visualization and analysis of protein interactions. *Bioinformatics* 19, 317-318.

Gagneur, J, Krause, R., Bouwmeester, T., and Casari, G. (2004). Modular decomposition of protein-protein interaction networks. *Genome Biol.* 5, R57. Epub.

Gardiner, K., Davisson M.T., and Crnic, L.S. (2004). Building protein interaction maps for Down's syndrome. *Brief Funct Genomic Proteomic* 3, 142-156.

Lu, H., Zhu, X., Liu, H., Skogerø, G., Zang, L., Zang, Y., Cai, L., Zhao, Y., Sun, S., Xu, J., Bu, D., and Chen, R. (2004). The interactome as a tree—an attempt to visualize the protein-protein interaction network in yeast. *Nucleic Acids Research* 32, 4804-4811.

Iragne, F., Nikolski, M., Mathieu, B., Auber, D., and Sherman D. (2004). ProViz: Protein Interaction Visualization and Exploration. *Bioinformatics* [Epub ahead of print]

Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41-42.

Kim, W., Bolser, D.M., and Park, J., (2004). Large scale co-evolution analysis of Protein Structural Interlogues using the global Protein Structural Interactome Map (PSIMAP). *Bioinformatics* 20, 1138-1150.

Kim, H., Park, J., and Han, K, (2003). Predicting Protein Interactions in Human by Homologous Interactions in Yeast. *Lecture Notes in Computer Science* 2637, 159-165.

Lappe, M., Park, J., Niggemann, O., and Holm, L. (2001). Generating protein interaction maps from incomplete data: application to Fold assignment. *Bioinformatics* Vol.17 Suppl.1, S149-S156.

Lehner, B. and Fraser, F. (2004). A first-draft human protein-interaction map. *Genome Biology* 5:R63 doi:10.1186/ gb-2004-5-9-r63.

Lawrence, M.C. and Colman, P.M. (1993). Shape complementarity at protein/protein interfaces. *J. Mol.*

Biol. 234, 946-950.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753.

Miller, S., Janin, J., Lesk, A.M., and Chotia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* 196, 641-656.

Moon, H.S., Bhak, J., Lee, K.H., and Lee, D. (2005). Architecture of Basic Building Blocks in Protein and Domain Structural Interaction Networks. *Bioinformatics* (in press).

Ng, S.K. and Tan, S.H. (2004). Discovering protein-protein interactions. *J. Bioinform Comput. Biol.* 1, 711-741.

Park, J., Dietmann, S., Heger, A., and Holm, L. (2000). Estimating the significance of sequence order in protein secondary structure prediction. *Bioinformatics* 16, 978-987.

Park, J. and Bolser, D., (2001). Conservation of protein interaction network in evolution. *Genome Informatics* 12, 135-140.

Park, J., Lappe, M., and Teichmann, S.A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* 30, 307, 929-938.

Poupon, A. (2004). Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr. Opio. Stru. Biol.* 14, 1-9.

Promislow, D.E. (2004). Protein networks, pleiotropy and the evolution of senescence. *Proc. R. Soc. Lond. B. Biol. Sci.* 271, 1225-1234.

Raicu V., Jansma D.B., Miller R.J., and Friesen J.D. (2004). Protein interaction quantified in vivo by spectrally resolved fluorescence resonance energy transfer. *Biochem J.* Sep 7;Pt. [Epub ahead of print]

Richard, F.M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* 82, 1-14.

Richards, F.M. (1977). Area, volumes, packing and protein structures, *Ann. Rev. Biophys. Bioeng.* 6, 151-176.

Varshney A., Brooks, F., and Richardson, D. (1995). Defining, Computing, and Visualizing Molecular Interfaces. *Proceedings IEEE Visualization* 95, 36-43

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Roder, L., Euzenat, J., Rechenmann, F., and Jacq, B. (1999). Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. *Nucleic Acids Res.* 27, 89-94.

Schof, J.W. (1993). Microfossils of the Early Archean Apex Chert: New Evidence of the Antiquity of life. *Science* 260, 640-646.

Searls, D.B. (1993). The Computational Linguistics of Biological Sequences. In *Artificial Intelligence and Molecular Biology*, L. Hunter, ed. (AAAI Press) pp.47-120.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403, 623-627.

Walhout, A.J., Boulton, S.J., and Vidal, M. (2000). Yeast two -hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* 17, 88-94.