14

Number Distribution of Transmembrane Helices in Prokaryote Genomes

Ryusuke Sawada and Shigeki Mitaku
Nagoya University
Japan

1. Introduction

Number distribution of transmembrane helices represents genetic feature of survival strategy, because the number of transmembrane helices is closely related to the functional group of membrane proteins: for example, most of membrane proteins that have six transmembrane helices belong to transporter functional group. Survival strategies were obtained by evolutionary mechanism that changes the genome sequences. Comparisons of number distributions of transmembrane helices among species that have different survival strategies help us to understand the evolutionary mechanism that has increased the categories of membrane proteins.

Some studies about how the categories of protein functions have been increased during evolution were performed using protein database (Chothia et al., 2003; Huynen & van Nimwegen, 1998; Koonin et al., 2002; Qian et al., 2001; Vogel et al., 2005). However, these studies were carried out by the analysis almost for soluble proteins. Classification of protein function groups are often carried out by the empirical methods such as sequence homology that use sequence information of three-dimensional structure resolved proteins as template sequences for each functional group. However three-dimensional structure resolved membrane proteins were much less than that for the soluble proteins because of experimental difficulty of membrane proteins.

In the previous study, we developed membrane protein prediction system SOSUI and signal peptide prediction system SOSUIsignal (Gomi et al., 2004; Hirokawa et al., 1998). By combination of those systems, number of transmembrane helices can be predicted based not on empirical but on physicochemical parameters. Therefore, it is possible to investigate the number distribution of transmembrane regions in membrane proteins comprehensively among various genomes by using SOSUI and SOSUIsignal.

2. Membrane protein prediction systems

SOSUI prediction software (Hirokawa et al., 1998; Mitaku et al., 2002) for transmembrane helix regions uses physicochemical features of transmembrane helix segments. Transmembrane helix regions have three common features: (1) a hydrophobic core at the center of the helix segment; (2) amphiphilicity at both termini of each helix region; and (3) length of transmembrane helix regions. These features are essential factors for the

transmembrane segment to stably present at the cell membrane. The SOSUI system first enumerates candidates of transmembrane regions by the average hydrophobicity of segments which are then discriminated by the distributions of the hydrophobicity and the amphiphilicity around the candidate segments.

SOSUIsignal (Gomi et al., 2004) predicts signal peptides that are removed from proteins that are secreted to the extracellular space via the secretary process. Signal peptides are present at the amino terminal segment of their respective proteins; the physicochemical features N-terminal structure is recognized by molecular modules during the cleavage process. The SOSUIsignal system is similar to the SOSUI system in that candidates are first enumerated by the average hydrophobicity at the amino terminal region and then real signal peptides are discriminated by several parameters.

By focusing on these physicochemical features, accuracy of the prediction systems is very good: approximately 95%for SOSUI and 90% for SOSUIsignal. By using these softwares, we can estimate not only function unknown protein sequence but also simulated ones.

3. Typical number distribution of transmembrane regions in membrane proteins

We investigated the population for number groups of transmembrane helices for 557 prokaryote genomes using SOSUI and SOSUIsignal. Figure 1 shows the results of the analysis of the membrane protein encoded in the *E. coli* genome as a typical example; the

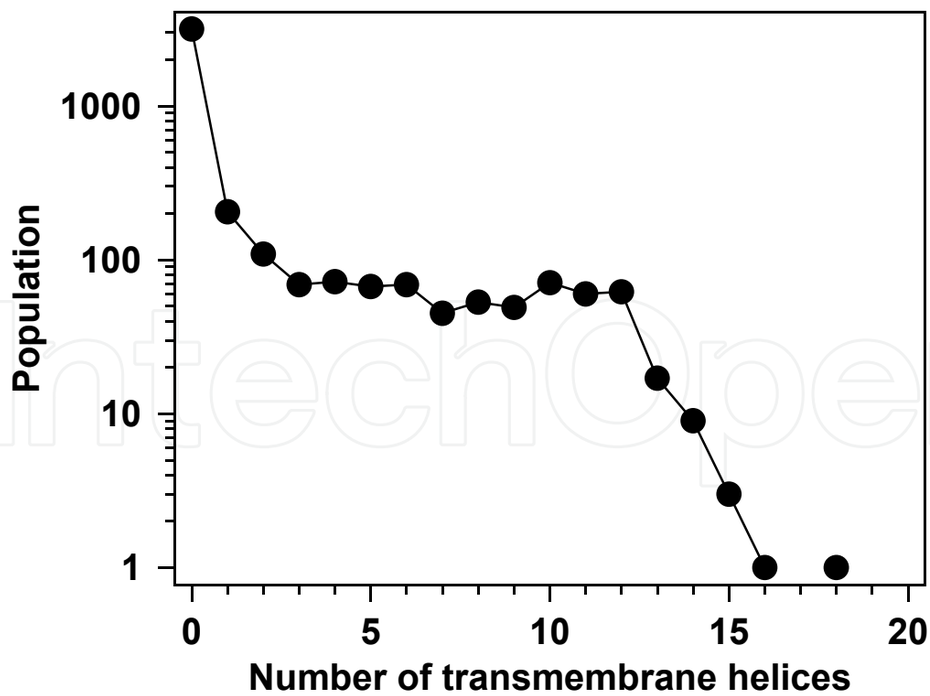


Fig. 1. Number distribution of transmembrane helices for *E. coli*. Estimation of the number of transmembrane helices for membrane proteins was performed using SOSUI and SOSUIsignal. Transmembrane helices number zero means soluble proteins.

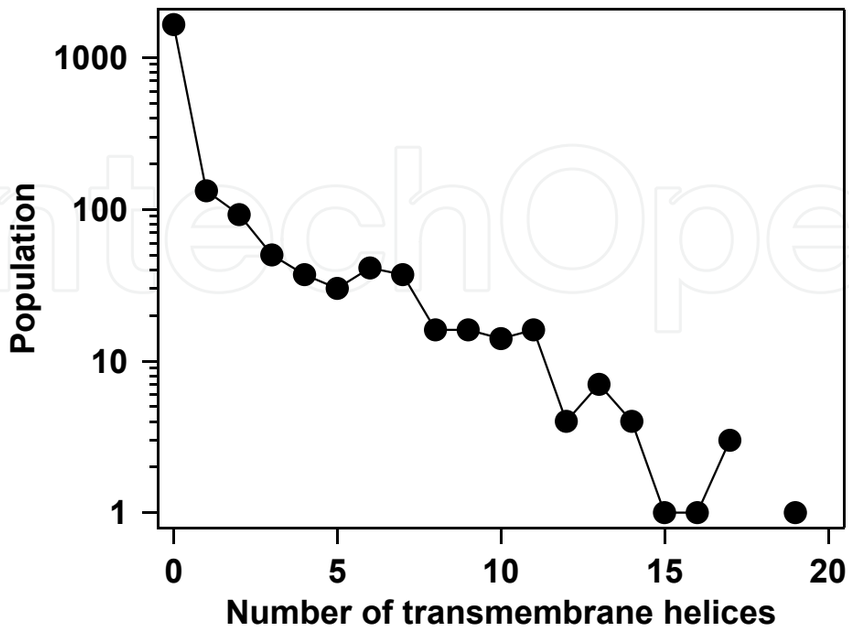
largest category of membrane proteins comprised proteins with only one transmembrane-spanning helix, and the second largest category comprised proteins with two transmembrane-spanning helices. The populations within each category decreased gradually up to 4 transmembrane helices and then there is a plateau from 4-13 helices**.** The population within each category decreased rapidly for categories comprising proteins with more than 13 transmembrane helices and there were apparently no proteins with more than 16 transmembrane helices. These results indicated that membrane proteins that have a particular number of transmembrane helices, such as 12, are important for *E. coli*.

4. Variety of number distribution of transmembrane regions among organisms

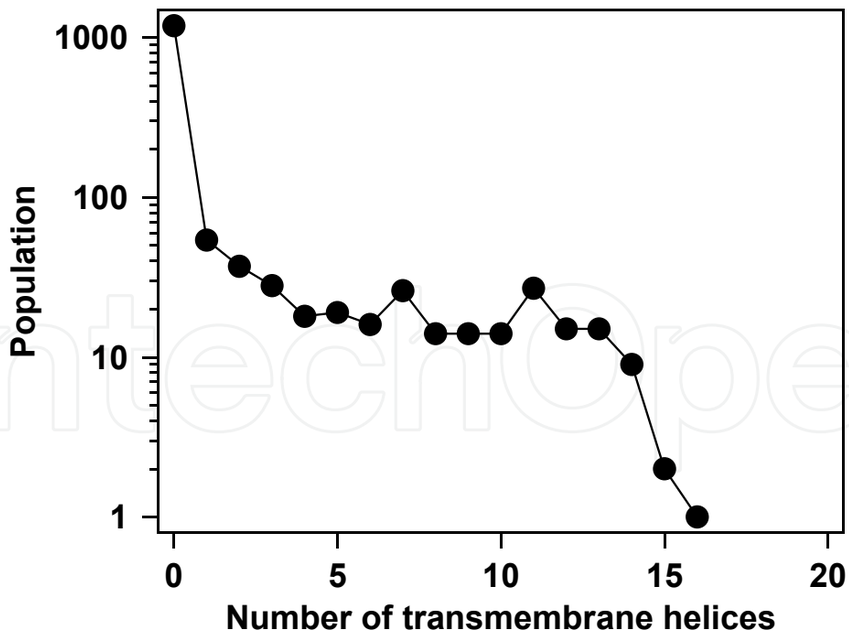
The general trend of the number distribution of transmembrane helices were very similar among 557 prokaryotic genomes, but the fine structures of the number distribution of transmembrane helices can change during the evolution. Four graphs in Fig.2 show the results for the analysis for four prokaryotic genomes: A, *Pyrobaculum calidifontis*; B, *Thermoplasma volcanium*; C, *Pseudomonas putida* and D, *Thermotoga petrophila*. We selected these four kinds of organisms for showing how the number distributions of transmembrane helices are different among organisms. The number distribution of transmembrane helices for *P. calidifontis* did not show significant shoulder at 13 helices as *E. Coli*. The shape of the distribution for *T. volcanium* was very similar to that for *E. Coli*, although the population was much smaller. A significant peak was observed at 12 helices for *P. putida*. A peak at 6 helices was observable for *T. petrophila*. Despite of the difference in the number distribution of transmembrane helices among organisms, the general trend of the distribution suggests the existence of a target distribution.

5. Number distribution of transmembrane helices in proteins in organisms grouped by GC contents

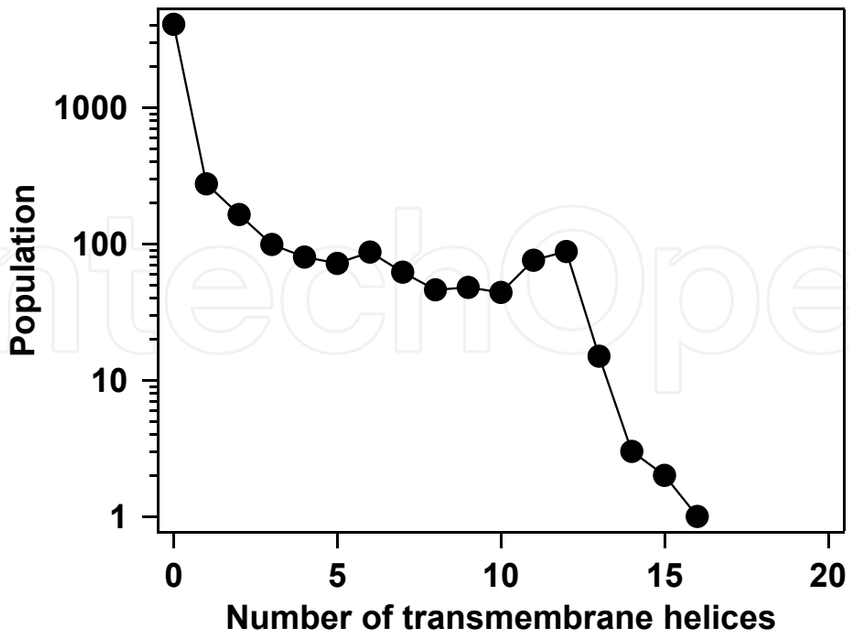
If the difference in the number distribution is due to the fluctuation around a target distribution, the difference would decrease by averaging of the distribution of many organisms. In contrast if the difference is due to some systematic change among organisms, the difference would not disappear by a simple procedure of the averaging. The GC content of genomes differs widely among species, from 0.3 to 0.7, and it is well known that various characteristics of prokaryotic cells systematically change according to GC content. Therefore, we investigated whether the distribution in the number of transmembrane helices per protein changed according to the GC content. Genomes for 557 prokaryotes were classified into nine groups with different GC content. In Fig. 3, the average number distributions of transmembrane helices in the nine groups indicated that the distributions were unchanged despite differences in GC content. The membrane-protein profile of the nine groups shared a common feature in that the general shapes of the curves were the same; the curves gradually decreased in the population of each category of membrane protein and there was a shoulder at the categories with 12 transmembrane helices. This result strongly suggests that the difference in the fine structures of the number distribution is due to the fluctuation around a general curve of Fig. 3. Then, aquestion arises about the natural selections: Is the general curve formed by the pressure of natural selection?


1000
100
10
1
Population
0
5
10
15
20
Number of transmembrane helices

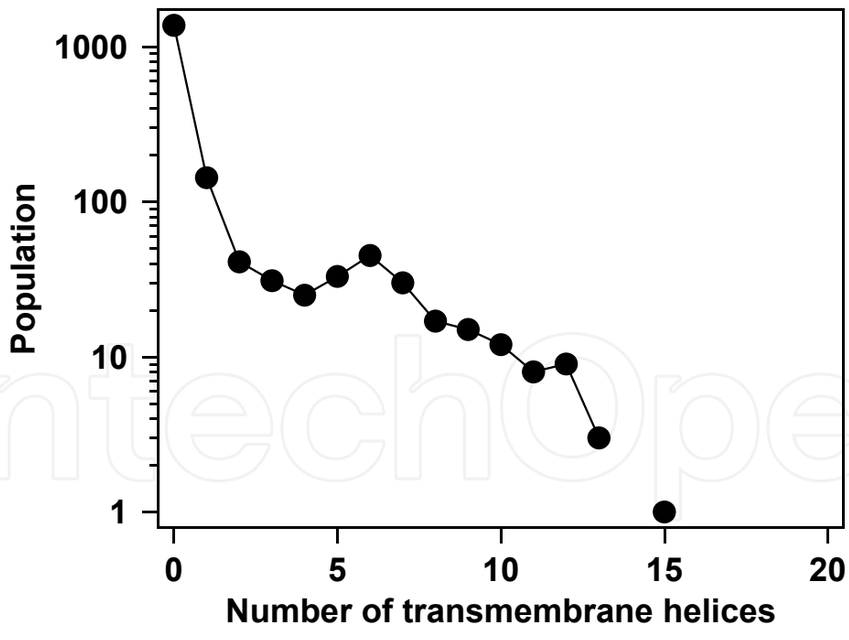
(a)


1000
100
10
1
Population
0
5
10
15
20
Number of transmembrane helices

(b)



(c)



(d)

Fig. 2. Number distributions of transmembrane helices for four prokaryotes *P. calidifontis* (A), *T. volcanium* (B), *P. putida* (C) and *T. petrophila* (D).

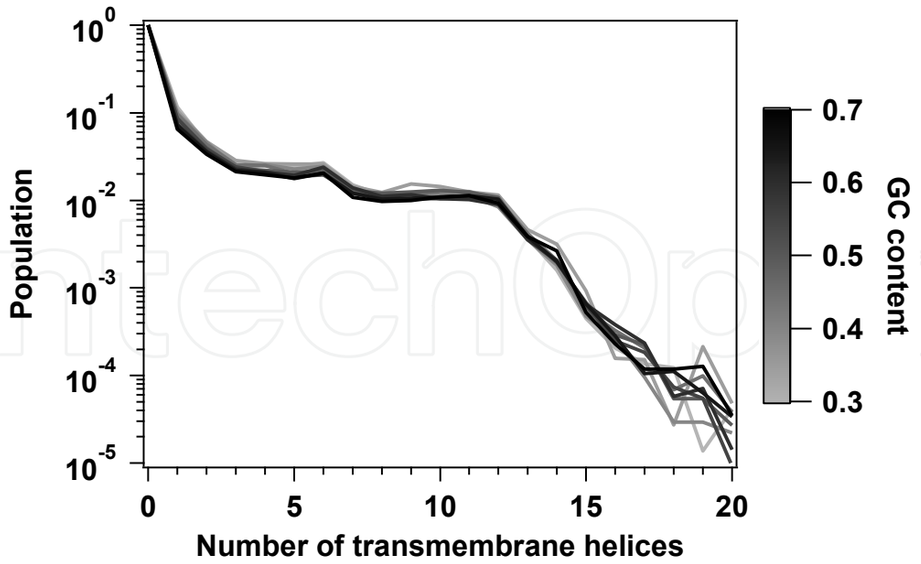


Fig. 3. Number distributions in nine groups of organisms classified by GC content. 557 prokaryotes were divided into nine groups according to GC content (0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65 and 0.7).

6. Random sequence simulation

Presumably, functionally important proteins are maintained in biological genomes by natural selection. Therefore, the general curve of the number distribution of transmembrane helices must reflect natural selection that occurs during biological evolution. The prediction systems, SOSUI and SOSUIsignal, have the great advantage that they are applicable to any amino acid sequences independent of empirical information because they are based mainly on the physicochemical parameters of amino acids. So, we planned to use the prediction system for comparing the number distributions of transmembrane helices between the real genomes and the simulated genomes in which comprehensive mutations are introduced with any pressure of the natural selection. Therefore, we investigated the effect of random mutation uncoupled from natural selection on the number distribution of transmembrane helices using random sequence simulations. The *E. coli* genome was used for the random sequence simulation. At each simulation step, one in every 100 amino acids in all protein sequences was mutated randomly. When the amino acids were mutated, the new amino acids were determined according to the genomic amino acid composition of the *E. coli* genome. Distributions of number of transmembrane helices were estimated by using membrane protein prediction systems SOSUI and SOSUIsignal after each simulation step. Simulations were reiterated until 500 simulation steps.

Distributions of transmembrane helices in membrane proteins for simulated genomes are shown in Fig. 4. As the simulation steps proceed, the number of membrane proteins with more than six transmembrane helices decreased monotonously and the shoulder in the distribution around 12 transmembrane helices disappeared. Beyond 300 simulation steps at which the sequences were completely randomized, the distribution became very similar to a single exponential decay. A broken line in Fig.4 represents the single exponential decay

curve, y = 2090e$^{-0.87x}$, which was obtained least square deviation analysis for the averaged distribution between 300 and 500 steps.

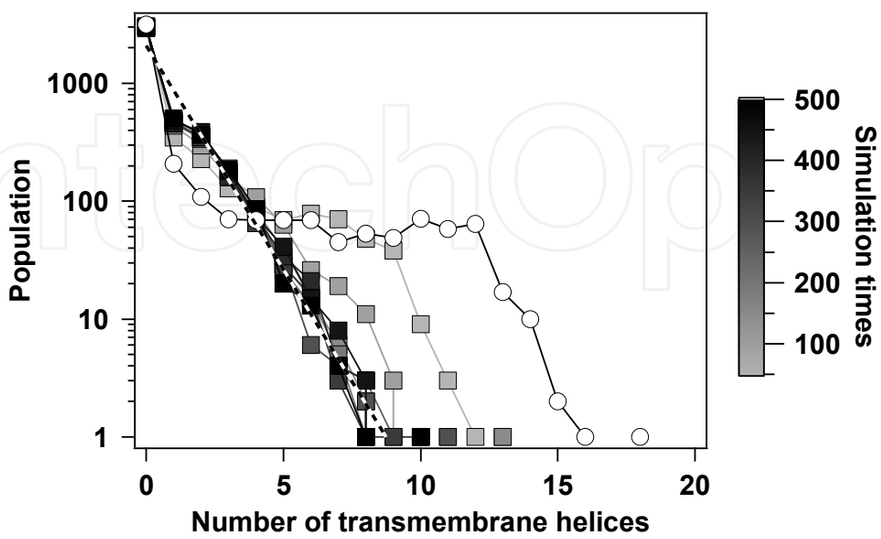


Fig. 4. Number distribution of transmembrane helices in *E. coli* genomes subjected to random mutation of amino acid sequences. Distributions of simulated genomes are represented as gray scaled rectangles. Open circle indicates the distribution for the original (simulation step zero) genome of *E. coli*. Dotted line represents the fitting result of y = 2090e$^{-0.87x}$ for the averaged distribution between 300 and 500 simulation steps.

After 300 simulation steps, shapes of the distributions of number of transmembrane helices for simulated genomes were almost unchanged in spite of additional mutations. A single exponential distribution for simulated genome can be explained by a kind of reaction in the evolutionary time scale changing the number of membrane proteins due to extensive mutations.

$$TM_i \leftrightarrow TM_{i+1}$$

in which TM_i represents a membrane protein with i transmembrane helices. If the equilibrium constant is the same among the distinct equilibrium state, the shape would become the exponential, as follow:

$$k^- \langle TM_n \rangle = k^+ \langle TM_{n-1} \rangle$$

$$\langle TM_n \rangle = \frac{k^+}{k^-} \langle TM_{n-1} \rangle = \left(\frac{k^+}{k^-} \right)^n \langle TM_0 \rangle$$

where <TM_n>, <TM_{n-1}> and <TM_0> represent the population of the membrane protein, with n, n-1 and 0 helices, respectively, and k^+/k^- means the equilibrium constant. In the simulation, the equilibrium constants for each transmembrane helices number group are the same from the algorithm of the prediction systems, and a single exponential decay in the computer experiment is well interpreted by this model. However, in the real genome, the shape of distribution is not exponential, showing a significant plateau and shoulder. This

indicates that there equilibrium constants for each transmembrane helices number groups are not same. This may be due to the difference of the functional importance among membrane protein groups.

7. References

Chothia, C., Gough, J., Vogel, C. & Teichmann, S.A. (2003). Evolution of the protein repertoire. *Science*, Vol. 300, No. 5626, 1701-1703

Gomi, M., Sonoyama, M. & Mitaku, S. (2004). High performance system for signal peptide prediction: SOSUIsignal. *Chem-Bio Informatics Journal*, Vol. 4, No. 4, 142-147

Hirokawa, T., Boon-Chieng, S. & Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, Vol. 14, No. 4, 378-379

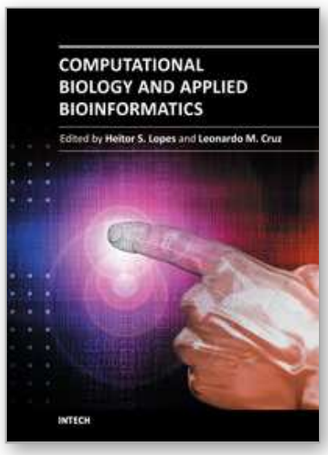
Huynen, M.A. & van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol*, Vol. 15, No. 5, 583-589

Koonin, E.V., Wolf, Y.I. & Karev, G.P. (2002). The structure of the protein universe and genome evolution. *Nature*, Vol. 420, No. 6912, 218-223

Mitaku, S., Hirokawa, T. & Tsuji, T. (2002). Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics*, Vol. 18, No. 4, 608-616

Qian, J., Luscombe, N.M. & Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol*, Vol. 313, No. 4, 673-681

Vogel, C., Teichmann, S.A. & Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. *J Mol Biol*, Vol. 346, No. 1, 355-365



Computational Biology and Applied Bioinformatics

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4

Hard cover, 442 pages

Publisher InTech

Published online 02, September, 2011

Published in print edition September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen