Bioinformatics as a problem of knowledge representation: applications to some aspects of immunoregulation

Sándor Pongor[1,2] and András Falus[3]

[1]*Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Area Science Park, I-34012 Trieste, Italy, e-mailpongor@icgeb.org*
[2]*Bioinformatics Group, Biological Research Center, Hungarian Academy of Sciences, Temesvári krt. 62, H-6726 Szeged, Hungary*
[3]*Deptartment of Genetics, Cell- and Immunobiology, Semmelweis University ,Immunogenomics Research Group, Hungarian Academy of Sciences, H-1089 Budapest, Nagyvárad tér 4, E-mail: faland@dgci.sote.hu*

Abstract

Bioinformatics uses a variety of models that fall into three broad categories such as linguistic, 3-D and interaction network models. Though latter allow one to capture interactions among molecules and other cellular components, the underlying representations are predominantly static. The main molecular mechanisms of immunology . such as VJD recombination, cellular and molecular networks, somatic hypermutations - cannot be and are not adequately covered in current molecular databases. Other aspects, such as the maturation of single, monospecific immune response or that of immunological memory apparently fall outside the scope of current molecular representations. The complexity of the immunological regulation such as polarized T cell cytokine web, Treg subpopulation, idiotypic networks, etc. calls for a new generation of computational approach leading to a new age of immunoinformatics ("immunomics").

1. Introduction

The growing network of biomedical databases and analysis programs constitute one of the most sophisticated knowledge representation tools mankind ever built. Bioinformatics differs from other informatics applications not so much by the amount of the data but rather by the complexity and the depth of knowledge it communicates. As an example,

bioinformatics deals with a wealth of molecular representations, such as sequences, 3/D structures, symbolic diagrams (e.g. hydrophobicity plots, helical wheel diagrams), as well as with a variety of group-wise representations such as multiple alignments, metabolic pathways, phylogenetic trees, etc., most of which could not have been conceived without computerized methods. It is customary to define bioinformatics as the informatics of biological data, but in fact it is not, or not exclusively a specialized branch of science: it is rather a general approach to all life sciences that makes it possible to study problems previously inaccessible to systematic research. This aspect – the access to new domains of knowledge – is one of the common themes that link the current age of computerized resources to previous innovations in storing and representing information. And as representations of information are at the very heart of cultural evolution, it is in place to introduce our subject within a historic context.

Complexity leaps in evolution are known to be powered by improvements in the way genetic information is stored and transmitted (Szathmáry and Smith, 1995, Maynard Smith and Szathmáry, 1995). By analogy one can point out that major improvement in scientific knowledge representation are correlated with innovations in the way information is shared within human societies, such as the appearance of writing, printing and the Internet. In a traditional society, knowledge is exchanged mainly by repeated, face-to-face communication and is confirmed and stored by an entire community. Writing not only decoupled knowledge transfer from personal communication but it also created a powerful new medium for the storage and manipulation of complex symbols whose interpretation required, at the same time, an increased intellectual effort from the recipient. Few would doubt that the widespread use of written and especially printed information has been a prerequisite of modern science that characterizes industrial societies. The paradigmatic knowledge source of this period is the *encyclopedia*, an organized, searchable knowledge base that is, in many respects, the predecessor of current electronic databases.

The current age of bioinformatics is characterized by vast amounts of biological data collected by computerized methods and distributed via the Internet and stored in electronic databases (**Table 1**). Knowledge in electronic databases is represented and transferred ways that is radically different from those known before. While readers can directly interpret printed text, electronic databases can only be “read” with the mediation of computer programs. Programs carry a large amount of implicit information in themselves that is not

always transparent to the user. For instance, in order to draw a three-dimensional picture of a protein molecule from an input of atomic coordinates, a program needs to know how the atoms of various types of amino acids are connected with each other. This kind of implicit information represents an intermediate layer between the data and the program and is often organized into *ontologies*, i.e. formal sets of definitions and rules that are valid for the data domain (**Table 2**). Also there are conspicuous changes in the way scientific information is confirmed. In the age of printed information, the quality of scientific discoveries was guaranteed by authoritative scientific societies, by the peer review of scientific journals, and last not least by the personal reputation of individual authors. In contrast, electronic databases are often produced by automated data collection, while textual annotations, such as the description of biological function, etc. are added by anonymous teams of database annotators who often rely on computer-based prediction methods. In other words, the amount of data and the number of databases is growing while data quality is less transparent. However, there are signs of integration as well. Large efforts are devoted to validating and interlinking biological data (sequences, structures), and, what is perhaps more important, highly complex scientific resources have been created wherein diverse data are controlled and accessed by uniform methods. In this setting, data integrity and quality will be increasingly controlled by autonomous agents, which will hopefully decrease the quality gap of current databases.

The first goal of this chapter is to provide of overview on how knowledge is represented in bioinformatics today and to show the cognitive roots of the underlying models. Bioinformatics is primarily concerned with the structure of protein and DNA molecules that fulfill functions in a series of interdependent systems such as pathways, cells, tissues, organs and organisms. This complex scenario can be best described with the concepts of systems theory[1]. Models in molecular biology are simplified systems that can be conveniently be described in terms of entities and relationships (Pongor, 1988). We will deal with three main kinds of representational models: language based models, 3-D models, and networks and will briefly review the development of computational tools that operate

[1] According to systems theory, a system is a group of interacting elements functioning as a whole and distinguishable from its environment by recognizable boundaries
CSÁNYI, V. (1989) *Evolutionary Systems and Society,* Durham and London, Duke University Press, KAMPIS, G. (1991) *Self-modifying systems in Biology and Cognitive Science,* Oxford, New York, Pergamon Press.. Molecules can be regarded as such systems. Generally speaking, structure is fixed state of a system, and the study of a system usually starts with its characteristic structures that are recurrent in space or time. Function on the other hand is not a property of the system rather a role that the system plays in the context of a higher system.

on these models. The last sections of this chapter describe genomic and immunological resources. The second goal of this chapter is to review the current knowledge on the information processing mechanism of the immune system. The immune system has specific algorithms for handling environmental information, and these mechanisms are among the most intensively researched and perhaps best understood phenomena in the life sciences that calls for specific informatics approaches yet to appear.

2. Sequences and Languages

Language-based descriptions are broadly speaking those that use semantic definitions for entities and relationships. The term "Molecular biology" was independently coined by Warren Weaver and John Astbury in the nineteen thirties, at a time when scientific methodology was dominated by linguistic theories (Wittgenstein, 1922, Carnap, 1939). Subsequent breakthroughs in information theory (Shannon, 1948b, Shannon, 1948a) and formal linguistics (Chomsky, 1957) all pointed towards a broad metaphoric context of language, communication and computation that provided the first framework within which genetics was discussed. Cryptography (Shannon, 1948c) and pattern recognition methods (Ripley, 1999), first developed within intelligence communities of the time, also contributed a great deal to the general view that biological sequences represent a code that carries information in a particular language, a metaphor reflected by such terms as the "genetic code" or "the book of life".

The analysis of biological sequences first used the statistical tools developed to analyze character strings in linguistics (Konopka, 1994), and many of the first methods such as those concerned with the string complexity, became standard bioinformatics tools in the later years. From the 1980's as bioinformatics became part of laboratory routine, pattern recognition methodologies that use similarity measures and classification algorithms proved to be of immediate interest, and with the onset of the genomic era, heuristic methods of searching biological databases such as BLAST (Altschul et al., 1990) became the most frequently used algorithms not only within bioinformatics, but allegedly in the entire field of scientific computing.

Margaret Dayhoff and her colleagues at the national Biomedical Research Foundation (NBRF), Washington, DC created first sequence database in the 1960s, an Atlas of protein sequences organized into families and superfamilies, and their collection center eventually became known as the PIR resource. Collections of DNA sequences (**Table1**), started at the European Molecular Biology Laboratory (EMBL, Heidelberg), at Los Alamos National Laboratory (New Mexico) and DDBJ, Japan, gained importance with the spreading of productive DNA sequencing technologies. Initially, sequence records included only the sequence the filename. These were eventually expanded to include annotation information such as references, function, regulatory sites, exons and introns, modified amino acids, protein domains etc. The Swiss-Prot collection of protein sequences is an especially good example of a well-annotated sequence database wherein a uniform syntax was developed for annotations. Very soon, separate so called secondary collections were created for annotated segments, such as the first protein domain sequence database (Pongor et al., 1993), as well as for posttranslational modifications, functional annotations, etc. (Table 4). Development of such secondary databases provided an important entry for specialized information, and current databases such as PFAM are excellent examples of this tendency. An important step was the application of WWW technology for cross-referencing the major databases with each other and subsequently with bibliographic database.

Information contained in current sequence databases(**Tables 3-4**) can be best pictured as an annotated sequence, a linear string of characters to which additional items of informations are linked either as an added field within the same record, or as a cross-reference to another database. Annotations items refer either to the entire sequence (global descriptors, such as protein name) or to a segment of it (local descriptors, such a protein domain or an exon). A database record itself can be an annotation items in a different database: for instance, a record in a protein domain database, or in a bibliographic database can be linked as an annotation items to a sequence database. In principle, annotated sequences should point to a unique protein or gene. In practice, a unique protein can be represented by many sequences, and databases differ in the way redundancy is handled. GenBank contains all published DNA sequences, so there is a considerable redundancy. The same is true for the protein sequence collections prepared by automatic translation, or by automated experimental procedures such as EST sequencing. Maintenance of high quality non-redundant databases require human overhead that is increasingly difficult to provide.

Finally, much of annotation information today is provided by automated procedures, such as similarity searches, HMM methods etc. Even though these methods constantly improve, there is no absolute guarantee behind the information, so much of annotation information today is labeled as "putative" or "by homology".

Interpretation of annotations and bibliographic records requires a uniform, computer-readable language. This need has fostered intensive research into the natural languages used in science. In early cultures, scientific language describing Nature cultures was based on only four elements (earth, water, fire, wind), and a few, dichotomic relations (hot-cold, dry-wet etc.) between them. Descriptions used in database annotations are based on a large number of models for atoms, molecules and reactions (**Figure 1**) This is a stripped-down language that can be kept uniform by the makers of the databases. On the other hand, scientific publications and abstracts use a "free-style" scientific language that is hard to handle by computers. This problem has emphasized the needs for developing common ontologies for molecular biology (Schulze-Kremer, 1997, Ashburner et al., 2000). An ontology defines a common vocabulary and a shared understanding within a domain of knowledge such as molecular biology. An ontology is an explicit description of the knowledge domain using concepts, properties and attributes of the concepts, and constraints on properties and attributes. The Gene Ontology Annotation (GOA) database (http://www.ebi.ac.uk/GOA) is a system of concepts and relations that is designed to convert UniProt annotation into a recognized computational format. GOA provides annotated entries for nearly 60,000 species and is the largest and most comprehensive open-source contributor of annotations to the GO Consortium annotation effort.

3. Three-dimensional models

As sequence databases were born within molecular biology, 3-D databases were brought to life by chemistry and later, structural biology. Same as language provides a conceptual framework for sequences, the metaphor for molecular models are common world objects whose handling and recognition is as at least as deeply rooted in human cognition as is language (Pinker, 2001). The first 3-D model of a molecule was constructed in 1874 by van t'Hoff who recognized that optical isomerism can only be explained by a 3-D arrangement of the chemical bonds. As methods of X-ray crystallography became applicable to organic molecules, Olga Kennard and Desmond. Bernal initiated the collection of 3-D structures what later became known as the Cambridge Crystal Structure Database, and access to 3-D information consolidated the use of molecular geometry in chemistry.

3-D descriptions of macromolecules are based on a series of concepts that resulted from several decades of scientific work. For instance structural descriptions of proteins is based on the stereochemistry of the peptide bond, but elements of secondary structures, supersecondary structures and finally protein folds are the joint results of several scientific disciplines that form what is called structural biology today.

The common ancestor of structural databases is the Protein Data Bank (PDB) (Berman et al., 2000), which was established in 1971 and six years later contained only 77 atomic co-ordinate entries for 47 macromolecules (Bernstein et al., 1977). Over the years, a conspicuous number of secondary databases have evolved from the PDB (see Table 2). Many of them concentrate on various classes of structural features, such as protein domains (Sander and Schneider, 1991, Orengo et al., 1997, Murzin et al., 1995, Siddiqui et al., 2001), loops (Donate et al., 1996), contact surfaces (Jones and Thornton, 1996, Luscombe et al., 2000), quaternary structure (Henrick and thornton, 1998), small-molecule ligands (Kleywegt and Jones, 1998), metals (Castagnetto et al., 2002) and disordered regions (Sim et al., 2001). Other databases concentrate on biological themes. The very concept of the "protein fold" owes much of its existence to such protein domain databases as CATH, SCOP and FSSP.

The conceptual structure of current 3-D databases is similar to sequence databases, inasmuch as the records contain both a structural description as well as an annotation part.

The definitions of the individual fields reflect the fact that PDB was originally created as a crystallographic database, and despite the fast growing body of NMR data, this remains its legacy. Current databases, such as ?? are now linked to other molecular and bibliographic databases.

Te development of ontologies has begun in structural biology. The STAR/mmCIF ontology (Westbrook and Bourne, 2000) of macromolecular structure is description of structural elements and data items in the framework of X-ray crystallographic experiments but it is extensible other kind of data collection techniques.

4. Genomes, proteomes, networks

Designing representations for genomes, proteomes and networks is a challenge as we deal with a wide variety of entities and relationships, partly predefined, partly discovered during the project. This class of representations can be called 'general topological model', wherein the nature of entities and relationships is not limited either to semantic or to 3-D concepts, as in the previous chapters. The resulting general representation is a graph wherein the physical entities are the nodes and their relations are the edges. The common ancestor of this representation is the structural formula, and graph theory itself owes a great deal to the development of chemistry in the nineteenth century. The representations of genomes as linear array of genes and other DNA segments follow a similar traditions tradition. The entities – genes – are predicted with gene-prediction programs or are determined experimental methods, and this adds a new layer of knowledge to the molecular data. The relationships are manifold but are predominantly binary in nature. Examples of relations include physical vicinity, distance along the chromosome, regulatory links extracted from DNA chip data and so on. The resulting picture is a graph of several ten thousand nodes and relatively few edges per node denoting various relationships. The description of proteomes is only somewhat different. The proteins are described in functional, biochemical and structural terms, and the relationships between proteins include metabolic relationships (sharing substrates in metabolic pathways) as well as structural relationships (sequence and structural similarities). Network models used in biology fall into two large categories. Dynamic models (such as metabolic network of a cell) are based on differential equations of the constituent the enzymatic reactions (for a recent review see:).Topological

models discussed here deal with the static properties of graphs, that can be undirected, directed and weighted.

From the computational point of view, genomes and proteomes are described as very large graphs in which the nodes (genes, proteins) and the edges (relations) are unknown or unsure. These large and fuzzy descriptions are in sharp contrast with the descriptions developed for well-defined graphs molecular structures, but the methods are not dissimilar to those used in other applications of graph theory. Given the large and varied genome sizes as well as the uncertainties of the data, genomic networks are usually characterized and compared in terms of gross global descriptors such as the degree distribution, composition (e.g. composition expressed in terms of gene- or protein classes). Biological networks are also believed to contain recurrent local patterns (network motifs) that are analogous to sequence motifs found in biological sequences.

Even this sketchy introduction implies that we deal with new a kind of complexity that originates, from the numerous and to a large extent, unknown interactions between the entities. On the other hand, the study of network topology in various fields – such as Internet, social- road- and electric networks, etc. – has provided interesting insights that have been successfully applied to genomes, proteomes and bibliographic networks. However these insight are limited by the fact that static network topology is only a very general description of the underlying biological phenomena, in fact it should rather be considered as a “metamodel” i.e. a “model of models”.

4. Computational tools

Bioinformatics came to life at a time when computer technology reached the daily routine of scientific research. The development of bioinformatics tools (e.g. interfaces, database design, programming methods) is to some extent a mere reflection of the concomitant trends in informatics. On the other hand, bioinformatics software is peculiar because of its impact on how lay users access biological data today. In the 1970's and early 80's, the first published programs were written in a basically sequential style for standalone computers such as campus mainframes. The second stage started by the recognition that the input and output of bioinformatics applications can be standardized, and modular packages based on the *software tools approach* (Knuth, 1998) were be developed. The best known example of these, the GCG package of John Devereux (Devereaux et al., 1984) developed into a battery of several hundred programs over the years, covering virtually the entire scope of biological sequence analysis. However, such a large body of knowledge is difficult to maintain in a commercial context. EMBOSS, which is developed by a collaboration of academic researchers was designed as an open source alternative commercial programs (Rice et al., 2000). The development of Bioperl (Stajich et al., 2002) – a *Perl* library for bioinformatics applications – and Bioconductor (Gentleman et al., 2004) – a statistical programming package for microarray analysis based on the *R* programming language – are further examples of successful open source collaborations. Web servers developed by academic research group represent a different trend since in such cases, the source code is often not released and users can access the programs only on-line. WWW technology thus allows individual researchers to release their programs before the commercial or open source development stage, and the users interested in the most recent computational tools more and more accept the risk of using non-transparent programs.

The most visible tools of current bioinformatics are the complex knowledge resources composed of databases, analysis tools and WWW interfaces that integrate various kinds of data into a navigable data network (**Figure 2**)

5. Information processing in the immune system

The immune response includes a number of regulating mechanisms that are organized into various networks affecting a wide range of phenomena ranging from uptake, processing and presentation of the antigens, to T and B cell activation and performing the effector functions. During the immune homeostasis, the spatial and temporal pattern of the cellular and soluble interaction networks develop the optimal qualitative and quantitative characteristics in starting, amplifying and finishing the immune response in an optimal way. The real understanding of immune response obviously requires a systems biology approach. There are four highly specific aspects of immune functions that warrant a specific immuno-informatic approach.

1. The immune systems itself functions as a highly regulated information processing system. The major informations are the sequence (αβTCR recognition) and the conformation (BCR/Ig and γδTCR recognition) of the antigen/peptide molecule, the genetic (e.g. MHC) background of the antigen presenting cells, the activation of the innate immune systems (e.g. complement, NK pattern), the actual environmental scenario (e.g. PAMP, local pattern of the inflammatory mediators, etc.).
2. The networking habits of the immune system, both at cellular level (enhancing and inhibitory effects, feed-back regulations) and as Ig networks, such as the idiotypic-antiidiotypic webs.
3. The VJD recombinations and other mechanisms generation diversity of the antigen receptor repertoires are basically different from other , more simple nets. These molecular events result in rapidly evolving gene sets serving a more sophisticated recognition tool during the afferent input of immune response.
4. Somatic hypermutations through *activation induced deamination (AID)* and repair machinery as highly effective way for generation of diversity belong to specific tools .

.In the following issues some relevant representations of immune regulation are mentioned

Information management by lymphocytes

What "program" the given cell has, namely what the "output" signal, the response is, issues from its genetic characteristics and the features it acquired during its ontogenesis

(Figure 3). In other words: after the appropriate co-stimulatory and cytokine effects, a cell-specific pattern of transcription factors develops. Upon their effect, the appropriate cell-response develops, thus, the cell divides or/and differentiates; its fate will be the transformation into a memory cell or apoptosis; it releases antibody, cytokines or other secretion products into the outer world. By this, it frequently participates in the activation or inhibition of another cell or functions as effector. Sometimes it happens that the cell "changes" its signal transduction way during the regulation. For instance, the cAMP-signal transduction related to MHC class 2 on B cells "switches" on to the tyrosine-kinase way.

The network of co-stimulatory effects

During the immune response, many cell-cell relations are formed and dissociated. Beyond the physical "approach", these cell-cell relations may cause intracellularly created signs, which change the functioning of the cell.

Besides the antigen-specific interaction (TCR-MHC/peptide), the T and B cells must receive supplementary signals for carrying out a successful immune response. Without these, the antigen-specific interaction in most case produces exactly produces anergy, incapability to respond. This anergy stands behind the phenomena of immune tolerance. Next to the specific antigen-receptor, there are co-receptors (CD4, CD8), whose binding to the appropriate MHC-molecules orienting T cells regarding which antigen-presenting cell (and exogenous or endogenous antigen) should be connected with.

Our present knowledge lists numerous co-stimulatory (formed with adhesion molecules and other membrane-proteins) interactions of positive and negative effect between the T cell and the antigen-presenting cell. Some interactions increase the activation, e.g. the CD28/B7.1, CD2-LFA-3 and CD40-CD40L (CD154) interactions (Sadra et al., 2004). The CTLA-4/B7.2 linkage, for instance, may have positive or negative effect depending on the circumstances (Zhang et al., 2003).

The possibility in relation with the special characteristics of NK-receptors raises a further regulation opportunity. The NKB1- and p58-receptors of NK-cells recognize MHC, but this recognition hinders the activity of NK-cells (KIR) (Wilson et al., 2000). It has been recently

stated that these *NK-receptors are also found on a subgroup of αβ–and γδ-cells*, where MHC-recognition is exactly a (positive) condition for the stimulus realized through the αβ–TCR-receptors. The point is that, within one cell, two receptor-structures (NK-receptor and TCR) are oppositely regulated by the MHC; thus, the outcome of cell-activation is influenced by the local proportion of the two kinds of receptors.

Presumably, the local, cell-level pattern of costimulating effects organized in time-order represent that fine regulation, which is optimal for starting, properly setting and concluding the immune response in the right time (Frauwirth and Thompson, 2002).

T-cell-dependent stimulation and inhibition, the Th1- and Th2-cytokines

Th-cells are heterogeneous considering their cytokine-production. It is proved that the cytokine-pattern of the Th1- and Th2-cells at the two ends of the polarized T-cell-lineage does not only deviate from one another, but, because of the cross-regulation, certain cytokines inhibit each other and each other's effect crosswise. This, together with the negative costimulation, has a significance in halting the response.

One of the most important effects of the IL-10 of Th2-origin is that it strongly inhibits the cytokine-production of Th1-cells, like, for instance, IL-2-synthesis (and its effects). The influence of IL-2 on B-cells is inhibited the same way by the IL-4, which also has mainly Th2-origin. The antagonism between the effects of IL-4 (Th2) and IFNγ (Th1) is extremely sharp and two-directional (e.g. on IgE-production, on DTH). Individual cytokines frequently affect the isotype of antibody production in different way, stimulating one and inhibiting the other (e.g. IgG-IgE). The cytokine-pattern determined by the tissue environment of B cells significantly influences the class-switch, that is, the development of the isotype (e.g. IgG or IgA antibodies are produced).

Usually it can be said (with exceptions) that T cells producing IFN-γ, IL-2, and TFN-β principally stimulate the cell-mediated immune response, while the T cells producing IL-4, IL-5, IL-9 and IL-6 mainly stimulate the humoral immune response. It is also interesting that chiefly B cells present the antigens to Th2-cells, while macrophages present the antigens to Th1-lymphocytes (Figure 4).

Elegant results have proved that the "Th1-Th2" character - related also to chemokine patterns, (Kim et al., 2001) - is not connected strictly to the CD4- marker, since this double-nature has been detected in some CD8+ cells as well. These cells are called Tc1/Tc2 cells. We also know Tγδ-1 which mainly produce IFN-γ and Tγδ-2 primarily secreting IL-4.

Th3 cells represent a separate subset characterized by TGF-β-production. They stimulate the functioning of Th1- and inhibit the functioning of Th2-subgroups. They also have important role in the IgA-production of the immune system attached to the gastro-intestinal system. Recently markers characteristic of the Th1- (LAG-3) and the Th2- (CD30) population have been found on the plasma membrane. LAG-3 is a molecule belonging to the immune-globulin supergene-family, while CD30 is a protein belonging to the TNF-receptor family, already known on activated T and B cells as well as on the Sternberg-Reed-cells typical of the Hodgkin-lymphoma.

If it is indeed true that these molecules are markers of the two T cell populations, the cytofluorometry which detects the surface markers, will be a supplementary method for measuring the Th1/Th2 rate besides the rather expensive cytokine-mRNA-measurements. As a result of T cell polarization, an important "division of labor" is formed during overcoming different infections. The role of the Th1-type Th cells is important against the Gram-negative bacteria, while the role of the Th2-type Th cells is essential against parasite infections (Figure 4.).

In these inhibiting and stimulating processes, the non-antigen-specific cells and products of the immune system also play an important role. For instance, if NK-cells are activated in the local immune reaction because of the antigen's nature (tumor-cell, virus) or other factors, this causes increase in the local IL-12 and IFNγ-level, that is, a "Th1"-like (+ and -) effect. According to present views, the IL-12 has a central role in the regulation of cellular (cell-mediated) immune response and promising results have come to light in the treatment of metastatic tumors and diseases caused by hepatitis B and C virus-infections, with the help of IL-12.

If the number of basophilic granulocytes increases locally because of the antigen's nature (allergen, vermin) or other factors, the result will be exactly opposite: it asserts the increase

of IL-4-level, that is, “Th2”-like influences (+ and -) (Figure 5.). On the other hand, IL-4 plays the “conductor’s” role regarding the humoral immune response. There is a therapeutic possibility in IL-10 (which is also supposed to have Th2-origin) in allergic diseases, because it inhibits the attractive effect of IL-5 on eosinophils.

Similar differences are caused by the locally effective prostaglandins produced by macrophages, fibroblasts and follicular dendritic cells. PGE1 and PGE2 inhibit the cytokine-production of Th1-cells but do not influence Th2-lymphocytes. As a consequence, PGE-s shift the balance locally into the direction of humoral immune response. We have also come to know (e.g. from the AIDS/HIV research) that corticosteroids principally inhibit Th1-cells (apoptosis-induction), while certain androgen steroids (e.g. dehydro-epi-androsteron) inhibit the corticosteroids’ Th1-blocking effect, that is, antagonize it. β–endorphin inhibits the rate of the Th1- and stimulates the rate of Th2-cells. β-antagonists weaken the cellular immune response through IL-12 inhibition.

Generally we can say that, in the organism, the Th1 and Th2 found on the two ends of the polarized T-cell-line, participate in different processes as regulating cells. Nevertheless, we should never explain the effects of the Th1/Th2 cytokines “dogmatically”, since the same cytokine can often act oppositely, depending on the concentration, place and time.

The CD4+NK.1.1 and Treg subgroup

Recent results have reported on a T cell type called CD4+NK.1.1+. It can be considered the main source of IL-4 and plays central role in anti-microbial immunity. The NK.1.1+ subgroup is a population which is CD4-CD8- αβ in the thymus, CD4+CD8- αβ on the periphery, having numerous NK-markers, cytotoxic ability and a relatively homogenous αβ-repertoire. Its function is supposedly the regulation of haematopoesis, the development of T cell tolerance, the cytotoxic removal of virus-infected liver-cells and the stimulation of the Th2 population’s maturation by IL-4. It is presumed that these cells primarily recognize microbial antigens presented by the monomorphic CD1 (Jiang and Chess, 2004).

On the basis of all these, the CD4+NK1.1 cells could be considered cellular elements of the non-antigen-specific immunity, a new (regulating?) subclass of the T cells (Godfrey and Kronenberg, 2004).

Most recently a new concept of Treg cells (CD4+, CD25+) develop (Walsh et al., 2004). These cells representing over 10% of CD4+ Th cells are mostly silencing cells expressing foxp3 transcription factor acting in various regulatory networks of immune response, producing TGFb, IL-10 and negative co-stimulatory molecules such as CTLA4 (Figure 6.). Treg cells are involved in transplantation tolerance, prevent pathological responses induced by the gut flora or microbial infections, play a role in maternal tolerance, can suppress antitumor immunity and protective immunity to pathogens and can lead to enhanced memory T cell response. Recently immunoregulatory disturbances of many autoimmune diseases (Frey et al., 2005) are coupled with dysfunction of Treg subsets.

Idiotype-regulation, idiotype network (Poljak, 1994)

The potential of antibody and TCR diversity developing in the immune response is extremely high. About 10^{11} different antibody- and 10^{15}-10^{18} different TCR-specificity develop in a healthy adult immune system. Since a significant part of this huge repertory is not expressed (or not in a significant degree) during maturation in the thymus, an auto-tolerance cannot be formed against them. As a consequence, the segments of the variable region (paratope, complexity-determining region-CDR) appear as antigens (idiotypes) in the organism. In 1973, Niels Jerne announced his attractive theory according to which the organism produces anti-idiotype antibodies. Antibodies (anti-anti-idiotypes) are formed again against the antigens in the variable regions of these antibodies and so on. This way, a network develops (Figure 7), where there is a possibility for every second element to contain similar epitopes (idiotopes) and have similar antibody-specificity. Thus, the first antibody activates a (B and Th cell) self-regulating network and the antibody playing the role of “antigen” activates the next element of “antibody” role. This, again activates the next one which may have “first antibody”-like idiotopes among its idiotopes. The network starts to limit itself; smaller and smaller amount of new antibodies are produced. Theoretically, the situation is similar in the variable regions of the TCR’s αβ– and γδ–chains. The idiotype-anti-idiotype network represents a complex network of interacting T and B cells, which can equally stimulate or inhibit the immunological activation. One of the

most important central principles of the network model is that the anti-idiotype antibody (second antibody) – which reacts with an idiotope in the CDR of another antibody (first antibody) – presents similarity with the original epitope, being the inner image of the original antigen-determinant. This theory is supported by facts; "second and third"-type anti-idiotype antibodies against monoclonal paraproteins have already been detected. By the improvement of detecting techniques researchers have recently identified anti-idiotype antibodies during normal immune response, as well. The idiotype network has a great significance in keeping the memory cells of the immune system in small but lasting excitement after the disappearance of the original antigen, in the presence of the second antigen, through an "internal image". This way, upon the repeated appearance of the "real" antigen, they can react quickly to it.

The proof showing the presence of the idiotype-network is the following: in a type of autoimmune thyroid disease, the anti-idiotype antibody acting against the autoantibody against the thyreoid-stimulating hormone (TSH) behaves like the TSH and binds to TSH-receptors. In the future, this phenomenon can be possibly used in vaccination. Anti-idiotype antibodies will be produced in an experimental animal against the antibody (specific for of using a quite dangerous, living pathogen) also produced in an experimental animal. This second antibody may be similar to original infectious antigen, therefore it can be used in immunization (securely, since it is an immunoglobulin), and thus, immunity develops in the organism.

Today debating opinions are published concerning the central role and size of a given idiotype network, but surely, this system has a considerable role in the regulation of the immune response. There are proofs about the existence of idiotype-specific Th-cells and idioype (antibody-, Th-) -cross reactions are also assigned importance in certain autoimmune diseases.

6. Concluding remarks

The general approach of bioinformatics emerged from a parallel growth in two major fields, life sciences and information technologies. This concomitant development provided access to several new fields, and it also resulted in new conceptual and technological tools for

representing and manipulating scientific knowledge. Integrated databases, analysis programs and ontologies are typical results of this development.

Current bioinformatics deals with a large number of models that are basically static in nature. Molecular database represent information in terms of linguistic, 3D and network models. Thowh latter allow one to capture part of the interactions among the molecules and other cellular componenets, the descriptions are still predominantly static. Large part of the information is stored in cellular, tissue and systemic models that are not part of the databases but – if they are represented at all - are stored in onthologies or are part of the background knowledged of biologists that is not accessible to computers.

The current databases of immunology are similar in structure and in philosophy to other molecular databases. Nevertheless immunology represents a specific case in many respects, and especially in the latter sense, since many of the models in immunology are different from those in other fields. This is on the one hand valid to the main molecular mechanisms of the immune system (such as VJD recombination, cellular and molecular networks, somatic hypermutations) which cannot be and are not adequately covered in current databases. On the other hand, major aspects of the immune system are not presently considered in molecular databases. As an example, the maturation of a single, monospecific immune response or that of immunological memory calls for a selectionist description of cell populations. Or, the similar structure-similar function paradigm that works magnificently at the level of most protein classes, is blatantly invalid in discriminating IgG molecules specific for different epitopes.

The overview presented in the present chapter suggests that the needs of immunological research will inevitably create a specific bioinformatics approach that will allow one to represent and access knowledge gained in these highly important fields.

Finally we mention that many biological concepts found their ways back to informatics: artificial neural networks and genetic algorithms are success stories of the computer sciences. It is an intriguing possibility that the information processing methods of immunology, the **immunomics concept** will inspire novel computational approaches (De Groot, 2004, Wang and Falus, 2004, Brusic and Petrovsky, 2005).

Table 1 Main types of bioinformatics databases[1]

Nucleotide Sequence Databases
- International Nucleotide Sequence Database Collaboration
- Coding and non-coding DNA
- Gene structure, introns and exons, splice sites
- Transcriptional regulator sites and transcription factors
- RNA sequence databases

Protein sequence databases
- General sequence databases
- Protein properties
- Protein localization and targeting
- Protein sequence motifs and active sites
- Protein domain databases; protein classification
- Databases of individual protein families

Structure Databases
- Small molecules
- Carbohydrates
- Nucleic acid structure
- Protein structure

Genomics Databases (non-vertebrate)
- Genome annotation terms, ontologies and nomenclature
- Taxonomy and identification
- General genomics databases
- Viral genome databases
- Prokaryotic genome databases
- Unicellular eukaryotes genome databases
- Fungal genome databases
- Invertebrate genome databases

Metabolic and Signaling Pathways

Enzymes and enzyme nomenclature

Intermolecular interactions and signaling pathways

Human and other Vertebrate Genomes
- Model organisms, comparative genomics
- Human genome databases, maps and viewers
- Human ORFs
- Human Genes and Diseases
- Model organisms, comparative genomics
- Human genome databases, maps and viewers
- Human ORFs

Microarray Data and other Gene Expression Databases

Proteomics Resources

Other Molecular Biology Databases
- Images of biological macromolecules
- Bioremediation database
- Drugs and drug design
- Molecular probes and primers
- Organelle databases

Plant databases
- General plant databases
- Arabidopsis thaliana
- Rice
- Other plants

Immunological databases

[1]Based on the Database issue of *Nucleic Acids Research*, 2005, http://www3.oup.co.uk/nar/database/cat/12/

Table 2 Databases for annotation terms, onthologies and nomenclature used in bioinformatics

Genew the Human Gene Nomenclature Database,
http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl
GO - Gene Ontology, http://www.geneontology.org/
GOA - Gene Ontology Annotation http://www.ebi.ac.uk/GOA
IUBMB Nomenclature database for enzymes http://www.chem.qmul.ac.uk/iubmb/
IUPAC Nomenclature database for organic and biochemistry, http://www.chem.qmul.ac.uk/iupac/
IUPHAR-RD Pharmacological nomenclature for receptors and drugs,
http://www.iuphar-db.org/iuphar-rd/
PANTHER Gene products nomenclature, http://panther.celera.com/
STAR/mmCIF: an ontology for macromolecular structure, http://ndbserver.rutgers.edu/mmcif UMLS - Unified Medical Language System (Thesaurus, lexicon and semantic networks)
http://umlsks.nlm.nih.gov

Table 3 Examples of protein sequence databases

Primary protein sequence resources	
Uniprot/Swiss-Prot - Annotated protein sequence db (University of Geneva, EBI)	http://www.expasy.org/sprot/
Uniprot/Trembl - Computer annotated protein sequences (EBI)	http://www.ebi.ac.uk/trembl/
Uniprot/PIR – Annotated protein sequences (Georgetown University)	http://pir.georgetown.edu/
Uniprot (Universal Protein database, Swissprot + PIR + TREMBL)	http://www.expasy.uniprot.org/
Secondary protein sequence resources	
COG - Clusters of Orthologous Groups of proteins	http://www.ncbi.nlm.nih.gov/COG
CDD – Conserved Domain Database	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
PMD - Protein Mutant db	http://pmd.ddbj.nig.ac.jp/
InterPro - Integrated Resources of Proteins Domains and Functional Sites	http://www.ebi.ac.uk/interpro/
PROSITE - PROSITE dictionary of protein sites and patterns	http://www.expasy.org/prosite/
BLOCKS - BLOCKS db	http://www.blocks.fhcrc.org/
Pfam - Protein families db (HMM derived) [Mirror at St. Louis (USA)]	http://www.sanger.ac.uk/Pfam/ http://genome.wustl.edu/Pfam/
PRINTS - Protein Motif fingerprint db	http://bioinf.man.ac.uk/dbbrowser/PRINTS/
ProDom - Protein domain db (Automatically generated)	http://protein.toulouse.inra.fr/prodom.html
PROTOMAP - Hierarchical classification of proteins	http://protomap.stanford.edu/
SBASE - SBASE domain db	http://www3.icgeb.trieste.it/~sbasesrv/
SMART - Simple Modular Architecture Research Tool	http://smart.embl-heidelberg.de/
TIGRFAMs - TIGR protein families db	http://www.tigr.org/TIGRFAMs/
BIND - Biomolecular Interaction Network db	http://www.bind.ca/
DIP - Db of Interacting Proteins	http://dip.doe-mbi.ucla.edu/
MINT - Molecular INTeractions	http://cbm.bio.uniroma2.it/mint/
ProNet - Protein-Protein interaction db	http://pronet.doubletwist.com/

Table 4 Examples of DNA sequence databases

Primary DNA sequence resources	
EMBL - EMBL Nucleotide sequence db (EBI)	http://www.ebi.ac.uk/embl/
Genbank - GenBank Nucleotide Sequence db (NCBI)	http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
DDBJ - DNA Data Bank of Japan	http://www.ddbj.nig.ac.jp/
dbEST - dbEST (Expressed Sequence Tags) db (NCBI)	http://www.ncbi.nlm.nih.gov/dbEST/
dbSTS - dbSTS (Sequence Tagged Sites) db (NCBI)	http://www.ncbi.nlm.nih.gov/dbSTS/
Secondary DNA sequence resources	
NDB - Nucleic Acid Databank (3D structures)	http://ndbserver.rutgers.edu/NDB/ndb.html
BNASDB - Nucleic acid structure db from University of Pune	http://202.41.70.55/www/net/deva.html
AsDb - Aberrant Splicing db	http://www.hgc.ims.u-tokyo.ac.jp/~knakai/asdb.html
ACUTS - Ancient conserved untranslated DNA sequences db	http://pbil.univ-lyon1.fr/acuts/ACUTS.html
Codon Usage Db	http://www.kazusa.or.jp/codon/
EPD - Eukaryotic Promoter db	http://www.epd.isb-sib.ch/
HOVERGEN - Homologous Vertebrate Genes db	http://pbil.univ-lyon1.fr/databases/hovergen.html
ISIS - Intron Sequence and Information System	http://www.introns.com/
RDP - Ribosomal db Project	http://rdp.cme.msu.edu/html/
gRNAs db - Guide RNA db	http://biosun.bio.tu-darmstadt.de/goringer/gRNA/gRNA.html
PLACE - Plant cis-acting regulatory DNA elements db	http://www.dna.affrc.go.jp/htdocs/PLACE/
PlantCARE - Plant cis-acting regulatory DNA elements db	http://sphinx.rug.ac.be:8080/PlantCARE/
sRNA db - Small RNA db	http://mbcr.bcm.tmc.edu/smallRNA/smallrna.html
ssu rRNA - Small ribosomal subunit db	http://rrna.uia.ac.be/rrna/ssu/
lsu rRNA - Large ribosomal subunit db	http://rrna.uia.ac.be/rrna/lsu/
5S rRNA - 5S ribosomal RNA db	http://rose.man.poznan.pl/5SData/
tmRNA Website	http://www.indiana.edu/~tmrna/
tmRDB - tmRNA dB	http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html
tRNA - tRNA compilation from the University of Bayreuth	http://www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/
uRNADB - uRNA db	http://psyche.uthct.edu/dbs/uRNADB/uRNADB.html
RNA editing - RNA editing site	http://www.lifesci.ucla.edu/RNA
RNAmod db - RNA modification db	http://medstat.med.utah.edu/RNAmods/
SOS-DGBD - Db of Drosophila DNA annotated with regulatory binding sites	http://gifts.univ-mrs.fr/SOS-DGDB/SOS-DGDB_home_page.html
TelDB - Multimedia Telomere Resource	http://www.genlink.wustl.edu/teldb/index.html
TRADAT - TRAnscription Databases and Analysis Tools	http://www.itba.mi.cnr.it/tradat/
Subviral RNA db - Small circular RNAs db (viroid and viroid-like)	http://nt.ars-grin.gov/subviral/
MPDB - Molecular probe db	http://www.biotech.ist.unige.it/interlab/mpdb.html
OPD - Oligonucleotide probe db	http://www.cme.msu.edu/OPD/

Table 5 Examples of three-dimensional databases

Primary 3-D resources	
Protein Data Bank	http://www.rcsb.org
Macromolecular Structure Database	http://www.ebi.ac.uk/msd/index.html
Nucleic Acid Database Project	http://ndbserver.rutgers.edu/NDB/index.html
BioMagResBank	http://www.bmrb.wisc.edu/Welcome.html
Protein domain/fold databases	
3Dee	http://jura.ebi.ac.uk:8080/3Dee/help/help_intro.html
CATH	http://www.biochem.ucl.ac.uk/bsm/cath
HSSP	http://www.sander.ebi.ac.uk/hssp
SCOP	http://scop.mrc-lmb.cam.uk.ac/scop
Examples of specialized resources	
BIND - binding database	http://www.bind.ca/index.phtml?page=databases
BindingDB – binding database	http://www.bindingdb.org/bind/index.jsp
Decoys 'R' Us	http://dd.stanford.edu
Disordered structures	http://bonsai.ims.u-tokyo.ac.jp/~klsim/database.html
DNA binding proteins	http://ndbserver.rutgers.edu/structure-finder/dnabind/
Intramolecular movements	http://molmovdb.mbb.yale.edu/MolMovDB/
Loops	http://www-cryst.bioc.cam.ac.uk/~sloop/Info.html
Membrane protein structures	http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html
Metal cations	http://metallo.scripps.edu/
P450 containing systems	http://www.icgeb.trieste.it/~p450srv/
Predicted protein models	http://guitar.rockefeller.edu/modbase
Protein-DNA contacts	http://www.biochem.ucl.ac.uk/bsm/DNA/server/
Protein-protein interfaces	http://www.biochem.ucl.ac.uk/bsm/PP/server/
ProTherm	http://www.rtc.riken.go.jp/jouhou/protherm/protherm.html
Quaternary structure	http://pqs.ebi.ac.uk
Small ligands	http://alpha2.bmc.uu.se/hicup/.
Small ligands	http://www.ebi.ac.uk/msd-srv/chempdb
The Protein Kinase Resource	http://pkr.sdsc.edu/html/index.shtml
Examples of search/retrieval facilities and database interfaces	
3DinSight – structure/function dbase	http://www.rtc.riken.go.jp/jouhou/3dinsight/3DinSight.html
BioMolQuest – structure/function dbase	http://bioinformatics.danforthcenter.org/yury/public/home.html
Entrez	http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi
Image Library of Macromolecules	http://www.imb-jena.de/IMAGE.html
OCA	http://bioinfo.weizmann.ac.il:8500/oca-docs/
PDBSUM	http://www.biochem.ucl.ac.uk/bsm/pdbsum/
ProNIT – protein/nucleic acid interactions	http://www.rtc.riken.go.jp/jouhou/pronit/pronit.html
TargetDB	http://targetdb.pdb.org/
SRS	http://srs.ebi.ac.uk/

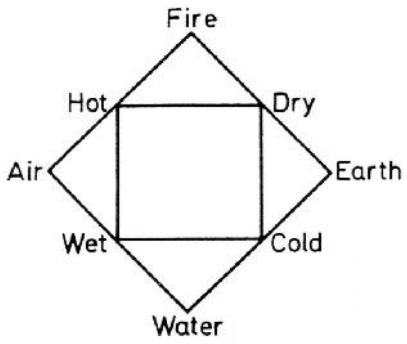
Table 4 Examples of genomic resources

Genomic databases for various organisms[1]	
Flybase - *Drosophila melanogaster*	http://flybase.bio.indiana.edu/
Subtilist - *Bacillus subtilis*	http://genolist.pasteur.fr/SubtiList/
Cyanobase - *Synechocystis* strain PCC6803	http://www.kazusa.or.jp/cyano/cyano.html
CYGD - *Sacharomyces cerevisiae*	http://mips.gsf.de/genre/proj/yeast/
ENSEMBLE – human and other invertebrate genomes	http://www.ensembl.org/
Comparative genomic visualization tools	
VISTA	http://www-gsd.lbl.gov/vista/
PipMaker	http://bio.cse.psu.edu/pipmaker/
Whole-genome annotation browsers	
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov
UCSC Genome Browser	http://genome.ucsc.edu/
Ensembl	http://www.ensembl.org/
Whole-genome comparative genomic browsers	
UCSC Genome Browser	http://genome.ucsc.edu/
VISTA Genome Browser	http://pipeline.lbl.gov/
PipMaker	http://bio.cse.psu.edu/genome/hummus/
Custom comparisons to whole genomes	
GenomeVista (AVID)	http://pipeline.lbl.gov/cgi-bin/GenomeVista
UCSC Genome Browser (BLAT)	http://genome.ucsc.edu/
ENSEMBL (SSAHA)	http://www.ensembl.org/
NCBI (BLAST)	http://www.ncbi.nlm.nih.gov/blast/

[1]A more complete list is available at the websites of the EBI, NCBI and DDBJ as well as within the current database issue of Nucleic Acids research, cited in Table 1.

Table 5 Bioinformatics resources for immunology

ALPSbase Autoimmune lymphoproliferative syndrome database	http://research.nhgri.nih.gov/alps/
BCIpep Experimentally determined B-cell epitopes of antigenic proteins	http://bioinformatics.uams.edu/mirror/bcipep/
dbMHC HLA sequences in human populations	http://www.ncbi.nlm.nih.gov/mhc/
FIMM Functional molecular immunology, T-cell response to disese-specific antigens.	http://research.i2r.a-star.edu.sg/fimm/
HaptenDB Hapten molecules	http://www.imtech.res.in/raghava/haptendb/
HLA Ligand/Motif database A database and search tool for HLA sequences	http://hlaligand.ouhsc.edu/
IL2Rgbase X-linked severe combined immunodeficiency mutations	*http://research.nhgri.nih.gov/scid/*
IMGT Integrated knowledge resource	http://imgt.cines.fr
IMGT-GENE-DB Genome database for immunoglobulins (IG) and T cell receptors (TR) genes from human and mouse	http://imgt.cines.fr/cgi-bin/GENElect.jv
IMGT/HLA HLA sequence database	http://www.ebi.ac.uk/imgt/hla/
IMGT/LIGM-DB Immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences, from human and other vertebrate species,	http://imgt.cines.fr/cgi-bin/IMGTlect.jv
Interferon Stimulated Gene Database (by microarry)	http://www.lerner.ccf.org/labs/williams/xchip-html.cgi
IPD-ESTDAB Polymorphic genes in the immune system	http://www.ebi.ac.uk/ipd/estdab/
IPD-HPA - Human Platelet Antigens	http://www.ebi.ac.uk/ipd/hpa/
IPD-KIR - Killer-cell Immunoglobulin-like Receptors	http://www.ebi.ac.uk/ipd/kir/
IPD-MHC Polymorphic genes in the immune system	http://www.ebi.ac.uk/ipd/mhc
JenPep Peptide binding to biomacromolecules within immunobiology (epitopes)	http://www.jenner.ac.uk/Jenpep
Kabat - Kabat db of sequences of proteins of immunological interest	http://immuno.bme.nwu.edu/
MHC-Peptide Interaction Database	http://surya.bic.nus.edu.sg/mpid
MHCBN Peptides binding to MHC or TAP	http://www.imtech.res.in/raghava/mhcbn/
MHCPEP MHC binding peptides	http://wehih.wehi.edu.au/mhcpep/
VBASE2 Gσerm-line V genes from the immunoglobulin loci of human and mouse	http://www.vbase2.org



System	Entities	Relationships
Molecules	Atoms	Atomic interactions (chemical bonds)
Assemblies	Proteins, DNA	Molecular contacts
Pathways	Enzymes	Chemical reactions (substrates/products)
Genetic networks	Genes	Co-regulation
Protein structure	Atoms	Chemical bonds
Simplified rotein structure	Secondary structures	Sequential and topological vicinity
Folds	C_α atoms	3-D vicinity
Protein sequence	Amino acid	Sequential vicinity

Figure 1. Examples of entities and relationships used in molecular models. Left: A simple fourfold semantics (earth, water, air, fire) and binary relations (hot-cold, wet/dry) was used in most early cultures as a conceptual framework to describe Nature. Right: Entities and relationships of modern databases.

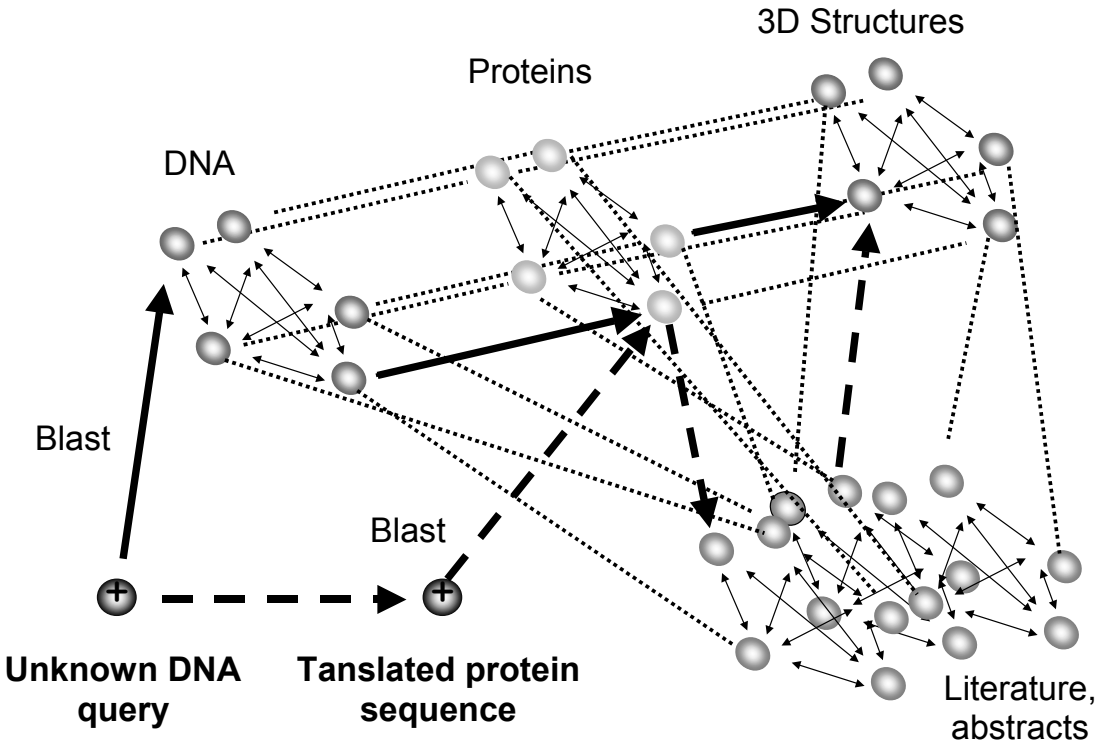


Figure 2. Search on an integrated database. Items in the individual databases (DNA sequence, Protein Sequence, 3-D structures, Literature abstracts) are cross-referenced (dotted line) by WWW links. Additional links (thin arrows) connect "neighbourhoods" i.e. similar data-items within each database. Consequently, if similarity search (thick arrows) points e.g. to an unannotated DNA entry, a member of its neighbourhood may help the user to find proteins, or protein structures.

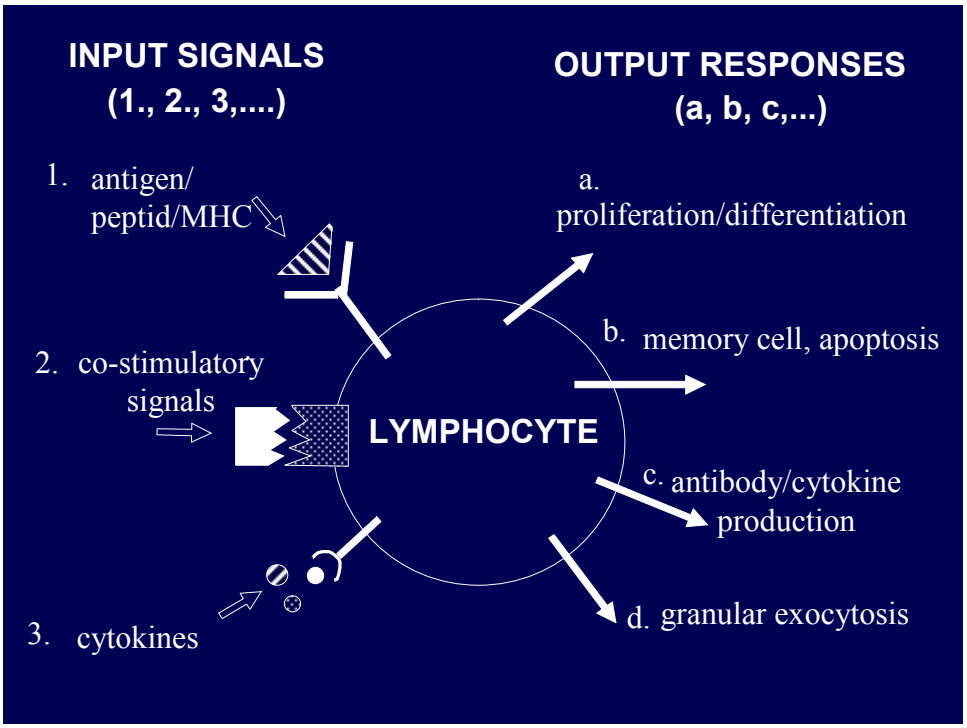


Figure 3. Information processing in lymphocytes.

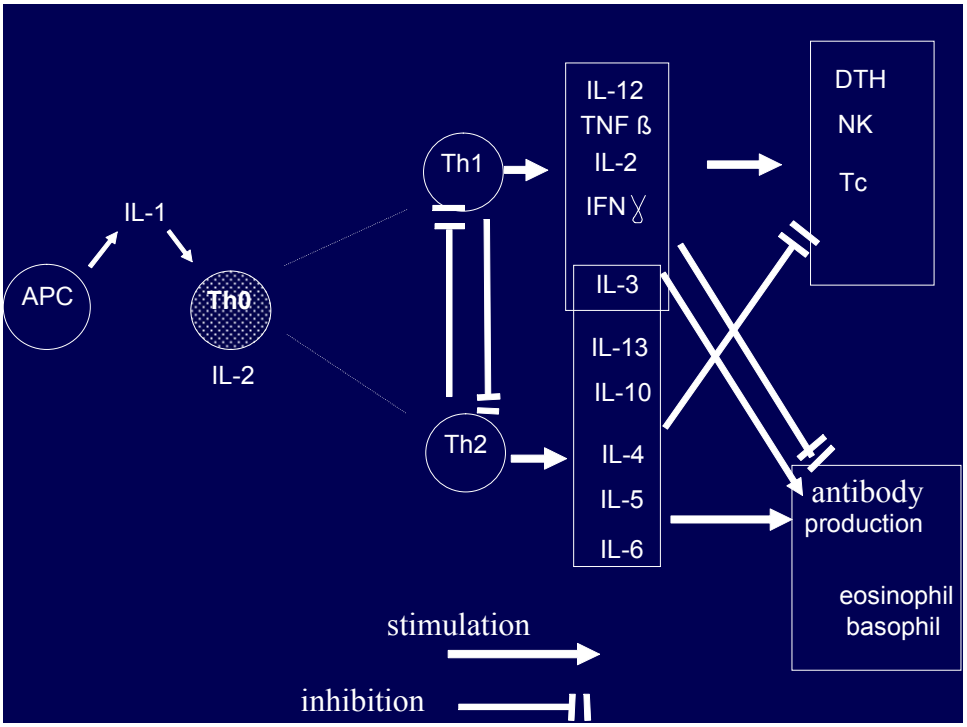


Figure 4. Cytokine patterns delivered by T cell polarization. Th1 cells tends to activate cell-mediated and suppress humoral immunity while Th2 cytokines have the opposite effect.

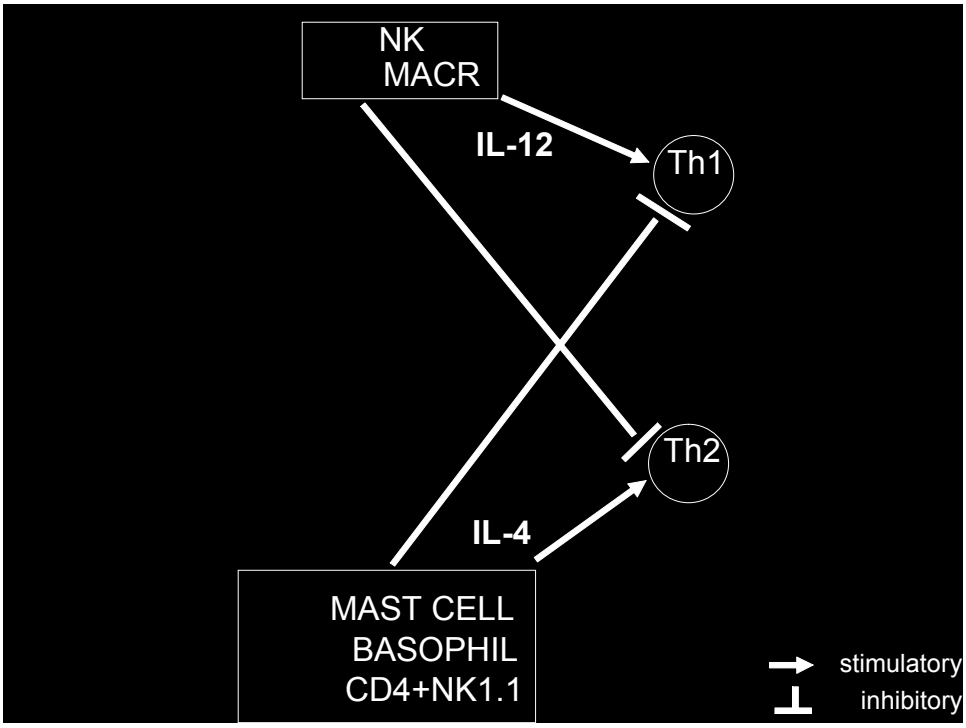


Figure 5. Innate immunity and T1/T2-like cytokine polarization. NK cells and macrophages produce preferentially "Th1 –like" cytokines, while mast cells, basophils and NK 1.1 cells secrete "Th2-like" cytokines

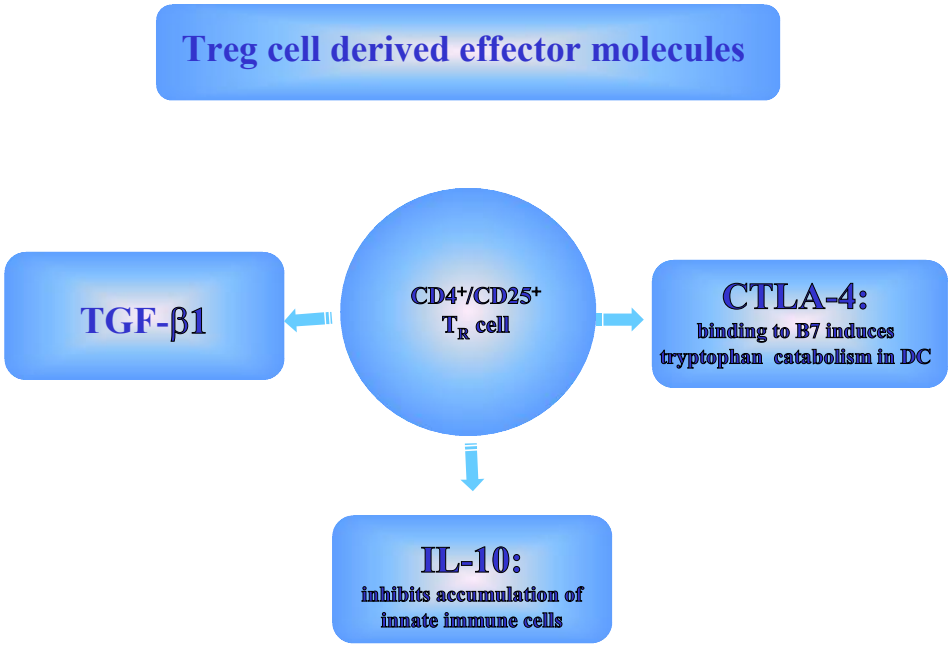


Figure 6. T regulatory cells expressing the negative costimulatory molecule CTLA-4 and secreting inhibitory cytokine TGF-β1 and IL-10.

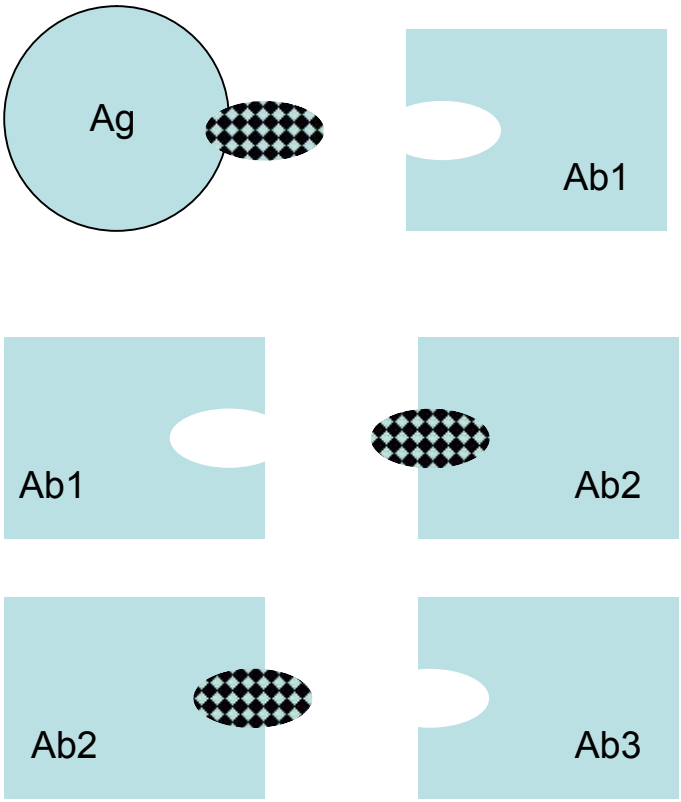


Figure 7. Idiotypic web. The antigen (Ag) is recognized by by a specific antibody (Ab1) carrying idiotope stimulating a second antibody (Ab2). Idiotypic determinant/s/ of Ab2 may further activate Abs, etc. Parts of antigenic structure may be closely similar to that of Ab2.

References

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28**,** 235-42.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J., MEYER, E. F., JR., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol,* 112**,** 535-42.

BRUSIC, V. PETROVSKY, N. (2005) Immunoinformatics and its relevance to understanding human immune disease. In press

CARNAP, R. (1939) *Foundations of Logics and Mathematics,* Chicago, University of Chicago Press.

CASTAGNETTO, J. M., HENNESSY, S. W., ROBERTS, V. A., GETZOFF, E. D., TAINER, J. A. & PIQUE, M. E. (2002) MDB: the metalloprotein database and browser at the Scripps Research Institute. *Nucleic Acids Res,* 30**,** 379-382.

CHOMSKY, N. (1957) *Syntactic Structures,* The Hague, Mouton.

CSÁNYI, V. (1989) *Evolutionary Systems and Society,* Durham and London, Duke University Press.

DE GROOT, A. S. (2004) Immunome derived vaccines. *Exp. Opin. Biol. Ther.,* 4**,** 767-772.

DEVEREAUX, J., HAEBERLI, P. & O, S. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research,* 12**,** 387-395.

DONATE, L. E., RUFINO, S. D., CANARD, L. H. & BLUNDELL, T. L. (1996) Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci,* 5**,** 2600-16.

FRAUWIRTH, K. A. & THOMPSON, C. B. (2002) Activation and inhibition of lymphocytes by costimulation. *J Clin Invest,* 109**,** 295-9.

FREY, O., PETROW, P. K., GAJDA, M., SIEGMUND, K., HUEHN, J., SCHEFFOLD, A., HAMANN, A., RADBRUCH, A. & BRAUER, R. (2005) The role of regulatory T cells in antigen-induced arthritis: aggravation of arthritis after depletion and amelioration after transfer of CD4+CD25+ T cells. *Arthritis Res Ther,* 7**,** R291-301.

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A. J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J. Y. & ZHANG, J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol,* 5**,** R80.

GODFREY, D. I. & KRONENBERG, M. (2004) Going both ways: immune regulation via CD1d-dependent NKT cells. *J Clin Invest,* 114**,** 1379-88.

HENRICK, K. & THORNTON, J. M. (1998) PQS: a protein quaternary structure resourse. *Trends Biochem Sci,* 23**,** 358-361.

JIANG, H. & CHESS, L. (2004) An integrated model of immunoregulation mediated by regulatory T cell subsets. *Adv Immunol,* 83**,** 253-88.

JONES, S. & THORNTON, J. M. (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A,* 93**,** 13-20.

KAMPIS, G. (1991) *Self-modifying systems in Biology and Cognitive Science,* Oxford, New York, Pergamon Press.

KIM, C. H., ROTT, L., KUNKEL, E. J., GENOVESE, M. C., ANDREW, D. P., WU, L. & BUTCHER, E. C. (2001) Rules of chemokine receptor association with T cell polarization in vivo. *J Clin Invest,* 108**,** 1331-9.

KLEYWEGT, G. J. & JONES, D. T. (1998) Databases in protein crystallography. *Acta Cryst. D,* 54**,** 119-1131.

KNUTH, D. E. (1998) *The Art of Computer Programming*, Addison-Wesley.

KONOPKA, A. K. (1994) Fundamentals of Biomolecular Cryptology. IN SMITH, D. W. (Ed.) *Biocomputing: Informatics and Genome Projects.* San Diego, New York, Boston, Academic Press.

LUSCOMBE, N. M., AUSTIN, S. E., BERMAN, H. M. & THORNTON, J. M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol,* 1.
MAYNARD SMITH, J. & SZATHMÁRY, E. (1995) *The Major Transitions in Evolution,* Oxford, New York, Heidelberg, W.H. Freeman.
MURZIN, A. G., BRENNER, S. E., HUBBARD, T. & CHOTHIA, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol,* 247**,** 536-40.
ORENGO, C. A., MICHIE, A. D., JONES, S., JONES, D. T., SWINDELLS, M. B. & THORNTON, J. M. (1997) CATH--a hierarchic classification of protein domain structures. *Structure,* 5**,** 1093-108.
PINKER, S. (2001) *The Language Instinct: How the Mind Creates Language*, Perennial.
POLJAK, R. J. (1994) An idiotope--anti-idiotope complex and the structural basis of molecular mimicking. *Proc Natl Acad Sci U S A.,* 91**,** 1599-1600.
PONGOR, S. (1988) Novel databases for molecular biology. *Nature,* 332**,** 24.
PONGOR, S., SKERL, V., CSERZÕ, M., HÁTSÁGI, Z., SIMON, G. & BEVILACQUA, V. (1993) The SBASE domain library: a collection of annotated protein segments. *Protein Eng,* 6**,** 391-5.
RICE, P., LONGDEN, I. & BLEASBY, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet,* 16**,** 276-7.
RIPLEY, B. D. (1999) *Pattern Recognition and Neural Networks,* Cambridge, Cambridge Univeristy Press.
SADRA, A., CINEK, T. & IMBODEN, J. B. (2004) Translocation of CD28 to lipid rafts and costimulation of IL-2. *Proc Natl Acad Sci U S A,* 101**,** 11422-7.
SANDER, C. & SCHNEIDER, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins,* 9**,** 56-68.
SCHULZE-KREMER, S. (1997) Adding semantics to genome databases: towards an ontology for molecular biology. *Proc Int Conf Intell Syst Mol Biol,* 5**,** 272-5.
SHANNON, C. E. (1948a) Communication theory of secrecy systems. *Bell Syst. Tech. J.,* 29.
SHANNON, C. E. (1948b) A mathematical theory of communication II. *Bell Syst. Tech. J.,* 27.
SHANNON, C. E. (1948c) Prediction and entropy of printed English. *Bell Syst. Tech. J.,* 30.
SIDDIQUI, A. S., DENGLER, U. & BARTON, G. J. (2001) 3Dee: A database of protein structural domains. *Bioinformatics,* 17**,** 200-201.
SIM, K. L., UCHIDA, T. & MIYANO, S. (2001) ProDDO: a database of disordered proteins from the Protein Data Bank (PDB). *Bioinformatics,* 17**,** 379-380.
STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A., DAGDIGIAN, C., FUELLEN, G., GILBERT, J. G., KORF, I., LAPP, H., LEHVASLAIHO, H., MATSALLA, C., MUNGALL, C. J., OSBORNE, B. I., POCOCK, M. R., SCHATTNER, P., SENGER, M., STEIN, L. D., STUPKA, E., WILKINSON, M. D. & BIRNEY, E. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res,* 12**,** 1611-8.
SZATHMÁRY, E. & SMITH, J. M. (1995) The major evolutionary transitions. *Nature,* 374**,** 227-32.
WALSH, P. T., TAYLOR, D. K. & TURKA, L. A. (2004) Tregs and transplantation tolerance. *J Clin Invest,* 114**,** 1398-1403.
WANG, E. & FALUS, A. (2004) Changing paradigm through a genome-based approach to clinical and basic immunology. *J Transl Med,* 2**,** 2.
WESTBROOK, J. D. & BOURNE, P. E. (2000) STAR/mmCIF: an ontology for macromolecular structure. *Bioinformatics,* 16**,** 159-68.
WILSON, M. J., TORKAR, M., HAUDE, A., MILNE, S., JONES, T., SHEER, D., BECK, S. & TROWSDALE, J. (2000) Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc Natl Acad Sci U S A,* 97**,** 4778-83.
WITTGENSTEIN, L. (1922) *Tractatus Logico Philosophicus,* London, Routledge and Kegan Paul.
ZHANG, X., SCHWARTZ, J. C., ALMO, S. C. & NATHENSON, S. G. (2003) Crystal structure of the receptor-binding domain of human B7-2: insights into organization and signaling. *Proc Natl Acad Sci U S A,* 100**,** 2586-91.