Monatsh Chem 139, 427–446 (2008)
DOI 10.1007/s00706-008-0892-6
Printed in The Netherlands

Monatshefte für Chemie
Chemical Monthly

Review

Modeling in biological chemistry. From biochemical kinetics to systems biology

Peter Schuster[1,2]

[1] Institute of Theoretical Chemistry, University of Vienna, Wien, Austria
[2] Santa Fe Institute, Santa Fe, NM, USA

Received 20 November 2007; Accepted 10 January 2008; Published online 14 March 2008

Abstract A brief review on biochemical kinetics in the twentieth century mainly concerned with enzyme kinetics and cooperative processes is presented. Molecular biology and, in particular, structural biology provided the basis for modeling biological phenomena at the molecular level. Structure was recognized as the ultimate and only level at which biological processes find an explanation that is satisfactory for chemists and physicists. A new epoch in biology was initiated by successful extensions of the molecular approach from individual molecules and reactions to the cellular and organismic level. Starting with sequencing of whole genomes in the 1980s more and more techniques became available that are suitable for upscaling from molecules to cells. A series of research programs was initiated: *genomics* dealing with sequencing the *DNA* of whole organisms, *proteomics* considering all proteins of a cell and their interactions, *metabolomics* studying all metabolic reactions of a cell or an organism, and *functional genomics* or *systems biology* aiming at an exploration of the dynamics of complete biological entities. At the same time computational facilities have experienced an unexpected development in speed of calculations and storing devices. At present computer simulations of whole cells at molecular resolution are within reach. The challenge for the theorist in biology is to develop methods for handling the enormously complex networks of gene regulation and metabolism in such a way that biological questions can be addressed. This goal cannot be achieved by dynamical systems theory alone. What is needed is a joint effort from different mathematical disciplines supported by empirical knowledge and tools from discrete mathematics to informatics. Two sections with selected examples from our own laboratory dealing with structural bioinformatics of *RNA* and with a dynamical systems approach to gene regulation are added.

Keywords Biochemical kinetics; Dynamical systems; *RNA* bioinformatics; *RNA* secondary structures; Systems biology.

Correspondence: Peter Schuster, Institute of Theoretical Chemistry, University of Vienna, Währinger Straße 17, A-1090 Wien, Austria. E-mail: pks@tbi.univie.ac.at

Chemical reactions, molecular structures, and cellular biology

In this section a historically motivated review of different mathematical techniques applied to problems in biochemistry and molecular biology is presented in three parts: (i) dynamical systems derived from chemical reaction kinetics, (ii) free energy optimization problems in predictions and design of biopolymer structures, and (iii) methods from discrete mathematics applied in the comparison and analysis of sequence data.

Biochemical kinetics

In the first half of the twentieth century mathematical modeling in biology was essentially based on the application of differential equations in two disciplines: (i) population genetics and (ii) biochemical kinetics. Other approaches were discrete models based on difference equations in discrete time intervals caused, for example, by seasonal cycles and discrete numbers of individuals or particles. The most popular discrete model goes back to the medieval mathematician *Fibonacci*[1]. Population genetics, founded by the three scholars *Ronald Fisher*, *J. B. S. Haldane*, and *Sewall Wright*, soon became a theory in its own right through uniting *Darwin*'s concept of natural selection and *Mendelian* genetics in the *Neo-Darwinian* theory of evolution. A large repertoire of analytical tools has been developed either by adopting methods from mathematics or by conceiving new techniques (see, *e.g.*, Ref. [1]). In particular, the models were extended to difference equations handling time in discrete intervals representing seasonal synchronization, to stochastic descriptions in order to account for phenomena related to small numbers of individuals in populations, and to random walk processes for situations were selection is absent [2]. Population genetics developed its own language that makes it sometimes hard to translate the results of molecular life sciences into this rather rigid conventional frame.

Biochemical kinetics branched of conventional chemical kinetics in the second decade of last century when *Michaelis* and *Menten* [3] published their seminal work on enzyme kinetics. Although *Michaelis-Menten* kinetics is neither a rigorous treatment of simple enzyme reactions nor a universally applicable approximation, it set the stage for more than ninety years of biochemical kinetics. Two new concepts based on experimental discoveries were decisive for the further development of modeling in biology: (i) methods for studying rapid reactions in solution, in particular stopped flow and relaxation techniques [4, 5] and (ii) statistical methods to study conformations and conformational changes of polymers [6], in particular models of biopolymers [7]. Rapid reaction techniques, in particular relaxation methods [8], enabled direct studies on elementary steps in biopolymer folding, conformational changes, and enzyme kinetics [9], and provided a new basis for the exploration of biochemical mechanisms. The theory of polymers provided also a first frame for studies on proteins and nucleic acids and yielded global statistical quantities that could be used as a reference for global characterization and comparison of different unfolded polymers. Originally derived from the *Ising* theory of ferromagnetism [10] the one-dimensional chain model for biopolymers [7, 11] provided the first approach towards a statistical mechanics of cooperative phenomena in the folding process of proteins or nucleic acids. Kinetics of cooperative transitions within the chain model was studied as well [11, 12]. In its simplest form the chain model uses only two thermodynamic parameters: (i) a nucleation parameter, σ, and (ii) a parameter for the local or single segment equilibrium, s. Numerous attempts were made to compute the two parameters from molecular data (for an early example see Refs. [13, 14], more recent work based on *Monte-Carlo* simulations is found in Ref. [15]). Many attempts were made to determine σ and s experimentally, as an example we mention here only a publication of more recent accurate measurements [16].

Biochemical kinetics has been in the focus of interest in the life sciences for three quarters of the twentieth century but then went through a period of stagnation between 1980 and 2000 before it regained importance in modeling gene regulation and metabolic reaction networks. At present it represents one of the major tools of computational systems biology [17] (see subsection on sequences, genetic information, and its processing and section on dynamical systems of genetic regulation networks).

Biopolymer structures

In the second half of the twentieth century new and extremely fruitful inputs into biology came from physics and chemistry: the techniques for the determination of molecular structures, primarily X-ray crystallography, were extended to investigations on biomolecules. The new discipline, structural biology, provided a straightforward explanation for the mech-

[1] *Leonardo Pisano* known as *Fibonacci* lived approximately 1170–1250 in Pisa and used the then already know progression, $a_{n+1} = a_n + a_{n-1}$ for $n \geq 3$ with the initial values $a_1 = 1$ and $a_2 = 1$ to model the growth of an isolated population of rabbit couples under the assumptions that (i) every month each couple gives birth to one new couple and (ii) the newborn couples start breeding after two months: $a_t = 1, 1, 2, 3, 5, 8, 13, \ldots$ for $t = 1, 2, 3, 4, 5, 6, 7 \ldots$ months.

anism of *DNA* replication, detailed insights into structures of proteins and other biomolecules, and allowed for molecular interpretations of protein functions[2]. A new paradigm for understanding and explaining mechanisms involving biological macromolecules was born:

$$\text{sequence} \Longrightarrow \text{structure} \Longrightarrow \text{function}$$

Crystallographic techniques for the determination of molecular structures of proteins and nucleic acids saw an impressive development, the resolution of X-ray diffraction methods has been increased to the level of atoms. In addition, several other techniques became available and the current repertoire of methods for elucidation of molecular structures comprises electron microscopy, various methods from molecular spectroscopy, in particular nuclear magnetic resonance, fluorescence, *Fourier* transform infrared and electron spectroscopy, as well as mass spectrometry. At present structure determination and interpretation of molecular properties and reactivities by means of known structures has become the standard of biological research. One more recent important step forward in understanding function in supramolecular complexes was the successful structure determination of the entire ribosome around the turn of the millennium [19, 20]: surprisingly, the catalytic molecule in the complex was found to be a *RNA* and not, as previously assumed, a protein molecule.

The accessibility of biopolymer sequences and structures provided a challenge for theorists: can structures be predicted from known sequences? If biopolymers were in a state of minimum free energy, the structure prediction problem would be tantamount to the search for the global minimum of a free energy surface in conformation space and structure prediction would boil down to an optimization problem. Apart from the occurrence of kinetically favored metastable states and specific interactions with other biomolecules in the environment *in vivo* small protein and nucleic acid molecules are mostly at thermodynamic equilibrium. For large molecules kinetic folding certainly has a strong influence on the native conformations. The question that is intimately related with the equilibrium hypothesis concerns the accessibility of molecular potentials or *conformational energy landscapes* of biopolymers [21]. Sufficiently accurate *ab initio* calculations are still not within reach and therefore one has to rely on empirically determined force fields [22, 23] or mean potentials, for example knowledge-based potentials from conformational ensembles of mean force [24–26].

The search for minimum free energy structures of molecules with known sequences, in essence, is an optimization problem that, however, was found to be notoriously difficult for proteins. Nevertheless, biopolymer structures are central to current biology and literature on prediction and design of protein and nucleic acid structures has become so extensive that we can mention here only a few typical examples. *De novo* protein folding[3] is not reliable enough, the alternative approach, comparative or homology modeling [27] suffers also from possible sources of errors related to the incompleteness of structure libraries with respect to natural protein folds [28]. An analysis performed five years later [29] comes to the much more optimistic conclusion that the current data bank is already sufficient for correct predictions, which are comparable to low-resolution experimental structures. The protein structure community performs regular contests in structure predictions, the last in the series was the ''Sixth Meeting on the Critical Assessment of Techniques for Protein Structure Prediction'' (CASP6) whose results are reported in Ref. [30]. The main conclusion is that only modest progress has been made over the last decade [31]. Structure prediction has a highly relevant inverse problem, the design of structures, which can be tailored for predefined purposes. Putting aside the structure-function relation part of the problem, *inverse folding* or *protein design* searches for sequences that fold into given structures (Fig. 1). In practice, protein design was first considered to be an extremely complex task and, indeed, the purely computational problem of protein design was proven to be NP-complete[4] [32]. A combined approach to protein design by theory and experiment, however,

[2] Instead of presenting individual references on the first investigations on biopolymers we recommend a monograph that describes the beginning and the first years of molecular and structural biology [18].

[3] Protein structure predictions that do not use direct input from known structures are called *de novo* folding.

[4] The notion NP-complete originated in computer science and is – somewhat sloppily – used for problems for which no algorithms exist that allow for finding solutions in polynomial time. *Polynomial time* means that the time required to find a solution increases with some power of the problem size.

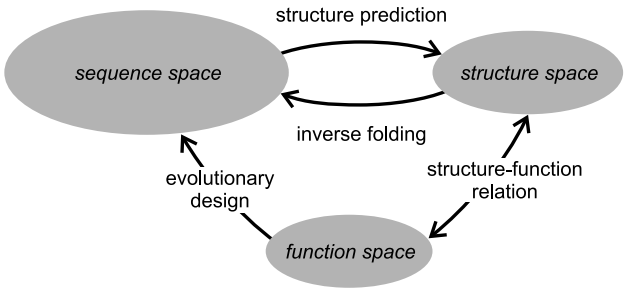

Fig. 1 Relations between sequences, structures, and functions in biopolymer design. The design problem is sketched in terms of mappings between three abstract spaces: sequence space, structure space, and function space. The sizes of the ellipses represent the current estimate of their cardinality. There are more sequences than structures and, presumably, there are more structures than functions (see the section on the prediction of structures and design of molecules)

turned out to be quite successful in many different applications [33–35]. Enormous scientific and commercial interest in *rational design* of proteins led to substantial progress within the last decade (see, *e.g.*, the special issue [36] and the recent reviews [37, 38]) and design software became a frequently used and indispensable tool in academia as well as industry.

Structures of nucleic acids fall into two categories: (i) antiparallel double helical and (ii) single strand structures. Double helical structures of *DNA* – with *Watson-Crick* base pairs exclusively – fall into different classes (A-*DNA*, B-*DNA*, *etc*.). They were long thought to be monotonous, and therefore boring. Molecular geneticists, however, were never satisfied with this idea. How could a regulatory protein then find its *DNA* target sequences with such high specificity? Crystallographic studies with model oligonucleotides beginning with the first high-resolution structure of a short *DNA* double-helix by *Dickerson* and *Drew* [39–41], however, showed high variability and pronounced sequence dependence of the B-*DNA* structure [42–44]. In addition, intermediate forms between A-*DNA* and B-*DNA* were found and a pathway based on crystal structures was suggested and interpreted in terms of sequence dependent stabilities of local conformations [45–48].

Structures of single strand nucleic acid molecules, predominantly *RNA*, turned out to be easier to handle than proteins, because there is a coarse-grained version of structure, the so-called secondary structure, which is much easier to handle and to predict. Secondary structures, in essence, are listings of *Watson-Crick* type base pairs formed in intramolecular antiparallel double helices. The formation of these substructures provides the largest stabilizing contribution to the free energy of folding. Minimum-free energy structures are computable by means of dynamic programming algorithms [49, 50]. *RNA* suboptimal structures and conformational partition functions are computable by dynamic programming algorithms as well [51–53], and folding kinetics can be modeled by means of an efficient stochastic process [54, 55]. Structure prediction is assisted by straightforward estimates on the reliability of the predicted structures. Moreover, algorithms for inverse folding [50, 56] became available. Computational nucleic acid design can be considered with different constraints like, for example, thermodynamic stability or efficient folding (see section on the prediction of structures and design of molecules). In addition, the relations between sequence, secondary structure, and function space in the sense of Fig. 1 are readily accessible in case of *RNA* secondary structures [57, 58].

Full three-dimensional structures of single strand nucleic acids are much harder to predict and to analyze. Different from protein spatial structures, however, there is a dominant interaction between side chains, namely base pair formation that is stronger than other interactions. Base pairs – *Watson-Crick* and others – can be classified in straightforward manner [59, 60]. Sequence specific recurrent motifs were found to dominate three-dimensional *RNA* structures [61, 62] and a recently formed *RNA* ontology consortium aims at completion of the *RNA* motif collection in order to provide enough information for the prediction of *RNA* structure from sequence [63].

An alternative approach to rational design mimics biological evolution in order to create molecules with predefined properties. In evolutionary biotechnology no structural information is required [64, 65] to design molecules with predefined functions (Fig. 1). Evolutionary design is straightforward and works particularly well with *RNA* molecules. An initial population with sufficient sequence diversity is created either through random synthesis or through replication with artificially high mutation rates. Population sizes up to $N \approx 10^{15}$ molecules have been used in typical test-tube evolution experiments. The desired function of the molecule is created through successive selection cycles (Fig. 2) consisting of (i) selection of the best suited molecules from a popu-

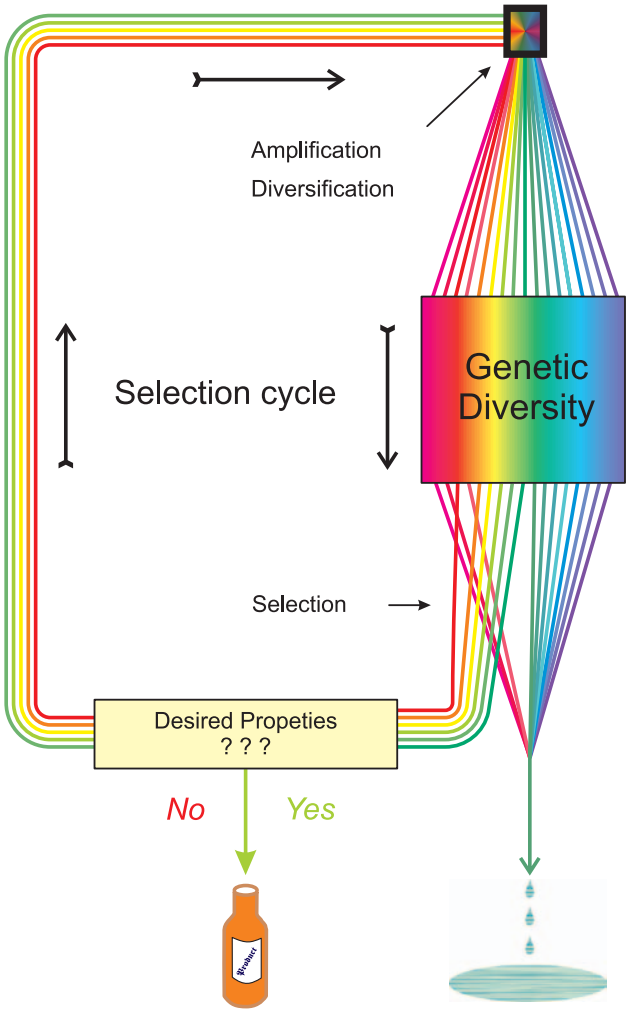

Fig. 2 The selection cycle in evolutionary design. Molecular properties are optimized in consecutive cycles consisting of selection from a sufficiently diverse pool of molecules, test for the desired function, amplification, and mutation of the selected molecules. The selection cycles are continued until molecules with the desired properties are obtained

lation, (ii) test of the desired function, (iii) amplification, and (iv) diversification through replication with adjusted mutation rates. The cycles are continued until molecules with desired properties are obtained. As an example for the selection step we mention the frequently applied technique of *systematic evolution of ligands by exponential enrichment* (SELEX) [66, 67]: a solution with *RNA* molecules is applied to a chromatographic column that contains covalently bound target molecules to which the *RNA*s are wanted to bind. Depending on the solvent all molecules are retained, which have stronger affinity to the target than some minimal binding constant. The retained molecules are eluted, amplified and diversified by mutation, and applied to the column again in a solvent that requires stronger binding to be retained. The procedure is continued until molecules with optimized binding properties are obtained. The kinetic theory of evolutionary optimization of *RNA* molecules in populations is well developed [68–70]. For investigations of stochastic phenomena computer algorithms are available for small populations up to $N = 100{,}000$ molecules [71–73].

Sequences, genetic information, and its processing

A novel era of molecular genetics began when large scale *DNA* sequencing became possible through the novel techniques of *Walter Gilbert* and *Frederick Sanger* [74, 75]. Automatic identification of labeled nucleotides [76] and computer assisted reconstruction of long *DNA* stretches and eventually of whole genomes [77] facilitated sequence data production, made sequencing substantially cheaper, and initiated the era of genomics. One of the major milestones towards the chemistry of life and the first highlight of the new sequencing approach to molecular genetics certainly was the determination and publication of the *DNA* consensus sequence of the human genome (International Human Genome Sequencing Consortium, 2001 [78]). Further improvement in *DNA* sequencing techniques is required when the dream of genome based personalized medicine should become true. This is not outside reach, since progress in single molecule techniques [79–83] has initiated new approaches towards single molecule *DNA* sequencing [84–88]. Genomics, in essence, is the successful upscaling of gene sequencing to the level of the entire cell. An at first surprising result was that only a minor fraction of about 5% of the human genome is used for coding proteins. For several years people thought that the rest of the genome – be it even 95% – was so called *junk* that remains unused as a remnant of phylogenetic history of man. Similar results were obtained for most higher organisms. The discovery of regulatory functions of *RNA* molecules – *small interfering RNA*s (si*RNA*s) and others – changed this common belief. In order to clarify the encoding function of *DNA* strong efforts were undertaken by a consortium staring the *ENCyclopedia Of DNA Elements* (*ENCODE*) project. The goal is to identify all *DNA* transcripts and to analyze their function. Recently, first results were reported for 30 million bases representing 1% of the human genome

[89]. One result is that ''the genome is by far more than a mere vehicle for genes'' [90] and *DNA* is pervasively transcribed, presumably more than 90% of the sequence appear at least in one transcript.

Comparison of *DNA* sequences through alignment and reconstruction of phylogenetic trees [91] required methods from discrete mathematics, in particular graph theory. Alignment of sequences started with dynamic programming algorithms based of two different scoring schemes: (i) the *Needleman-Wunsch* scoring scheme [92] leading to the best global alignment of two sequences and (ii) the *Smith-Waterman* scoring scheme [93] returning the best local alignment. Since then an enormous variety of different techniques for the alignment of two and more sequences have been developed and now fast and reliable software is available for large scale comparisons of sequences and for the search of databases (see, *e.g.*, Ref. [94]). The availability of whole genomes for the reconstruction of the *tree of life* revealed a number of surprises. In particular, it turned out that horizontal gene transfer is much more common in prokaryotic life than expected [95–97] and this might well jeopardize the existence of a tree of life during early precambrian development and render futile the search for such a tree [98–102]. More recent estimates based on data from more genomes [103, 104] show, however, that this fear was exaggerating the effect of gene migration between species and kinds of consensus trees do nevertheless exist. A challenge for theorists, nevertheless, remains: new methods of sequence comparisons are required that do not presuppose the existence of a single unique tree [105–107].

In case of *RNA* viruses the structure of the *RNA* genome encodes not only proteins but also the life cycle of the phage [108]. Molecular structures of viral *RNA*s are often conserved despite substantial sequence variation. The same is true for *RNA* molecules with functions based on their structures. Then, simultaneous sequence alignment and structure prediction becomes a highly relevant issue. An algorithm developed by *Sankoff* [109] solves the problem in principle. For applications to natural *RNA* molecules, however, the *Sankoff* algorithm is not suitable, since it requires $O(n^4)$ memory and $O(n^6)$ CPU time. All practical implementations employ heuristics to reduce the search space. The first such attempt was Foldalign by *Gorodkin et al.* [110] which allowed only simple stem-loops. Meanwhile, many other simplified versions are available which reduce search-space by restricting possible sequence alignments or possible structures, or both [111–120]. Other notable approaches derive the conserved structure without sequence alignment [121, 122].

The *DNA* sequence, in essence, is the chemical formula of a *DNA* molecule, and contains the encoded information for the cellular synthesis of proteins and *RNA* molecules and as such provides no direct access to their spatial structures nor does it tell the function of the biomolecules. The next logical step is to investigate the translation products of genes and to develop methods to discover their interactions [123]. Precisely this is the goal of proteomics[5] [124]. Application of chip technology [125] and mass spectrometry [126] opened previously unknown possibilities. This new field developed novel techniques to analyze proteins, to study interactions between proteins, the two-hybrid systems [127, 128] for example, and to perform high throughput investigations.

The next step in the development of modern genetics brings biochemical kinetics back in the focus of interest. Indeed, systems biology and/or quantitative biology aim at quantitative description and modeling of reaction dynamics in entire cells and organisms (for a comprehensive review of different techniques applied to genetic regulation and metabolic networks see Ref. [129]). The current computer work on dynamics is almost entirely dealing with integration of kinetic equations derived from biochemical reaction networks. Because of the enormously large number of molecular species in cellular reaction networks, modules of metabolism rather than whole cells are frequently studied. One of the first studies of this kind was dealing with glycolysis in yeast [130]. A special language, s*ystems biology markup language* (SBML), has been developed [131] in order to allow for an automated formulation of kinetic differential equations from input data describing the reaction mechanism. Appropriately SBML is directly combined with an efficient ODE solver, for example CVODE [132].

For real biological systems the problem of upscaling dynamical systems to the enormously high dimensions of metabolic networks is still unsolved. Several kinds of more or less suitable approxima-

[5] As the genome or the genotype is genetic information upscaled to the cellular level, the *proteom* is the set of all proteins and their interactions in the cell.

tions have been applied. One example is the replacements of kinetic equations by piecewise linear ODE in properly defined segments of concentration space [133]. Other approaches perform metabolic flux analysis [134–139] or consider the topology of cellular networks by statistical methods [140–143].

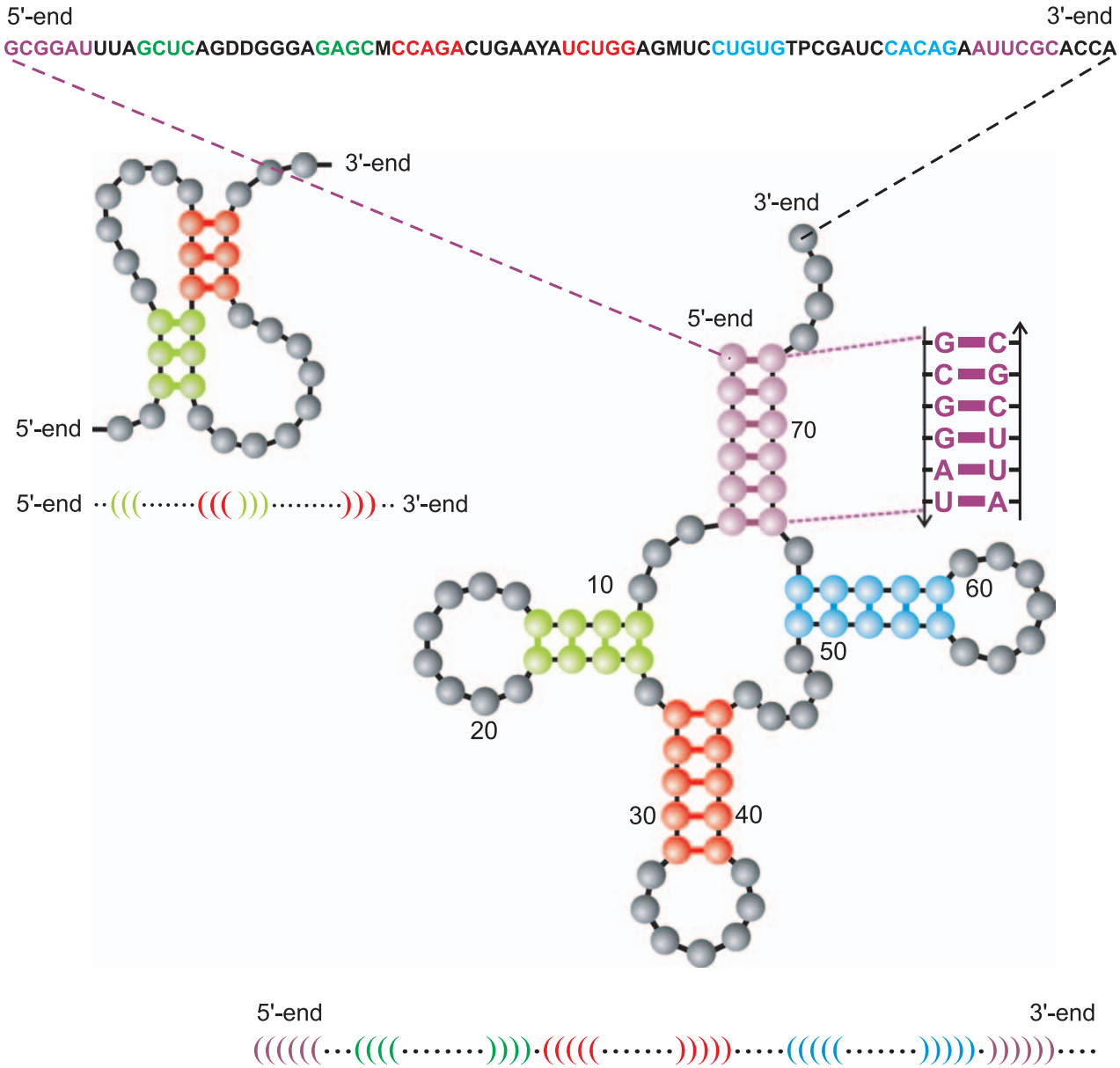

Fig. 3 *RNA* secondary structures. The figures show the sequence, the conventional graph of the secondary structure, and its symbolic notation for phenylalanyl transfer *RNA*. The principle of folding *RNA* sequences into secondary structures consists of double helix formation with *Watson-Crick* and **G–U** base pairs under the condition of free energy minimization. The four double helical stacking regions are indicated in color. As shown in the insert on the r.h.s. the backbone strand folds back on itself leading to antiparallel orientation in the double helix. In the symbolic notation at the bottom of the figure, which is equivalent to the secondary structure graph in the middle, base pairs are represented by parentheses and single nucleotides are shown as dots. Coloring of base pairs is dispensable since the left and right parts of parentheses are assigned to each other by conventional mathematical notation requiring the absence of pseudoknots in the secondary structure (see, *e.g.*, Ref. [58]). The sequence contains a number of modified nucleotides (**D** = dihydro-uracil, **M** = methyl-guanine, **Y** = wyosine, **T** = thymine, **P** = pseudouridine), which among other properties have the effect to stabilize the cloverleaf structure of the t*RNA*. The insert on the l.h.s. shows a structure with a (H-type) pseudoknot. In this case the assignment of base pairs requires colors

Prediction of structures and design of molecules

Secondary structures of single stranded *RNA* molecules provide a basis for the mathematical analysis of the relations between sequences and structures. The secondary structures, in essence, are listings of *Watson-Crick* and **G–U** base pairs in antiparallel double helical regions, which are formed through folding the backbone strand back on itself (Fig. 3). A commonly applied condition for *RNA* secondary structures is the absence of so called pseudoknots[6]. The main motivation for the neglect of pseudoknots is technical: computation of minimum free energy structures with pseudoknots is enormously more time consuming than the standard algorithms [144]. A justification for the approach, however, can be seen in the fact that pseudoknots are rare and they can be introduced into optimized secondary structures as tertiary interactions.

The number of possible pairing patterns for a given sequence is very large and increases exponentially with chain length n. All these pairing patterns can be understood as suboptimal conformations of one *RNA* molecule. Most suboptimal conformations have positive free energies of formation[7]. It is straightforward to neglect suboptimal conformations with positive free energies unless they are required as intermediates along lowest passes of lowest free energies from one conformation to another. The remaining number of suboptimal conformations is still very high. The set of suboptimal conformations for a given sequence can be computed by means of algorithms: *Zuker* [52] computes all conformations of most but not all classes, *Wuchty et al.* [51] compute all suboptimal conformations within a defined energy band above the minimum free energy, and the algorithm developed by *John McCaskill* [53] calculates the partition function based on the secondary structures of all suboptimal conformations.

[6] A pseudoknot is defined as an *RNA* structural element containing conventional base pairs of (approximate) *Watson-Crick* geometry, **G**≡**C**, **A**=**U**, and **G**–**U**, and unpaired nucleotides, which when written in the symbolic notation requires colored parentheses for uniqueness. In other words, the symbolic notation violates mathematical parentheses assignment (Fig. 3).

[7] The energy of formation is computed as the free energy difference between the structure under consideration and the unfolded or random coil chain. A positive free energy of formation implies that the conformation is unstable in comparison with the random coil chain.

Structure prediction

Computation of minimum free energy (mfe) secondary structures is commonly performed by means of algorithms based on dynamic programming [49, 145, 146]. Additivity of the free energies of structural elements is assumed. These structural elements are loops (hairpin loops, internal loops, bulges, and multi loops), flexible elements (joints or free ends), and stacks of two adjacent base pairs. Loops and flexible elements consist of unpaired nucleotides and provide a destabilizing mainly entropic contribution to the free energy (for a comprehensive review see Ref. [58]). Structure stabilizing contributions come from base pair stacking in the double helical regions. Free energies, energies, and entropies of the structural elements are introduced as empirical parameters and are derived from kinetic and thermodynamic data measured on model compounds, which are commonly either synthetic or natural oligo-ribonucleotides [147, 148]. Empirical data are also available for folding of single-stranded *DNA* [149, 150]. The usage of integer algebra allows for substantial speedup of the computations [50]. A major problem for the reliability of *RNA* secondary structure prediction is the existence for low lying suboptimal conformations that can erroneously become the mfe structure because of limited accuracy of parameters and approximate nature of the assumptions, for example the additivity of the free energies of substructures. Another source of errors are tertiary interactions, which by definition are not included in secondary structure computations (the different classes of pseudoknots are just one example). Tertiary interactions may change the base paring pattern on the secondary structure in the minimum free energy conformation and cause errors in this way.

Structures formed from unfolded sequences under laboratory conditions or in nature need not coincide with the minimum free energy structure. Kinetic effects may determine the observed conformation. An illustrative example is the hairpin/double-hairpin switch shown in Fig. 4. The double-hairpin is formed approximately twice as often as the single hairpin, because it has two nucleation sites compared to one in the hairpin, and accordingly the distribution of structures is 33/67 whereas the equilibrium mixture would favor the hairpin in a ratio of 88/12. A substantial difference between the kinetically preferred and the thermodynamically determined distribution

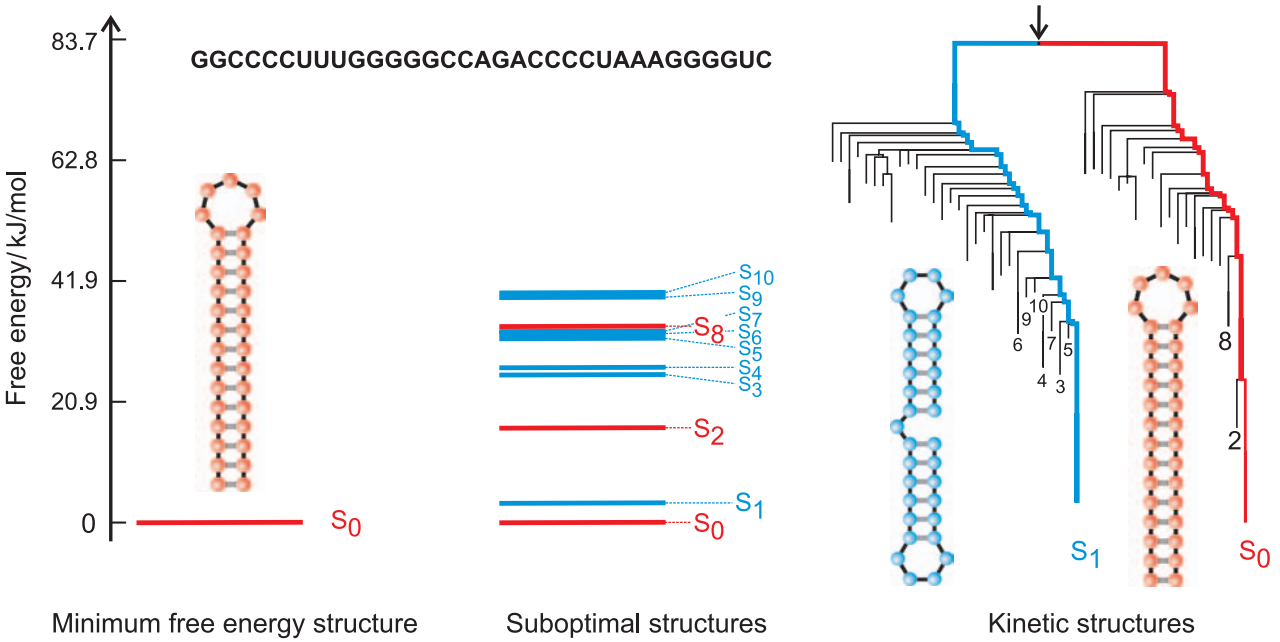

Fig. 4 *RNA* structures and suboptimal conformations. The figure sketches three commonly used notions of *RNA* secondary structure for an *RNA* molecule of chain length $n = 33$ with the sequence shown in the insert. The minimum free energy (mfe) structure (l.h.s.) is obtained by conventional folding algorithms based on dynamic programming computing the structure of minimum free energy for a given sequence (mfe: $\Delta G^0 = -110\,\mathrm{kJ/mol}$ relative to the unfolded sequence). In the middle we show energies of the mfe structure and the suboptimal conformations. The interconvertibility of conformations through saddle points is shown by means of the barrier tree (r.h.s.): individual suboptimal conformations are connected *via* the paths of lowest free energies, which are represented only by the free energies of the two local minima and the lowest saddle connecting them. The example shown in the figure represents a so-called *RNA* switch, an *RNA* molecule that can exist in two dominant conformations denoted by S_0 (red, being the mfe structure) and S_1 (blue, being the lowest suboptimal conformation with a free folding energy of $\Delta G^0 = -105\,\mathrm{kJ/mol}$), S_2 and S_8 (*red*) and higher lying states are suboptimal conformations in the basin of S_0, whereas S_3, S_4, S_5, S_6, S_7, S_9, and S_{10} (*blue*) and others belong to S_1. Free energies of energy levels are given relative to the mfe

of conformations is to be expected for large *RNA* molecules [151]. Moreover, synthesis by transcription starts always from one end and secondary structure formation goes on during synthesis. Refolding yielding the thermodynamically favored substructures becomes unlikely when the sequences are sufficiently long and a partial secondary structure has been formed that is stable enough against unfolding at the temperature of the experiment.

Suboptimal conformations are computed straightforwardly by means of dynamical programming: the energy table is computed as in the case of mfe structure calculation and all conformations are obtained by extending backtracking to all possible paths. The only problem is the enormously large number of suboptimal states even when they are restricted to negative free energies. Possible ways out of the dilemma are the neglect of certain classes of conformations [52] or the restriction to conformations within a predefined energy band above mfe [51]. The free energies of the entire set of suboptimal structures can be visualized as a free energy landscape provided an appropriate notion of distance between conformations is introduced, which is compatible with the move set for kinetic folding [58]. The whole spectrum of suboptimal structures shows families of conformations that share common structural features. These families are related to basins of the conformational free energy landscape to be discussed below.

Folding kinetics at the resolution of single base pairs can be formulated as a stochastic process and simulated by means of trajectory calculations and trajectory sampling [54]. A move set for folding kinetics is defined in such a way that every conformation can be reached from every conformation, for example base pair cleavage and base pair formation. For economic computational performance it turned

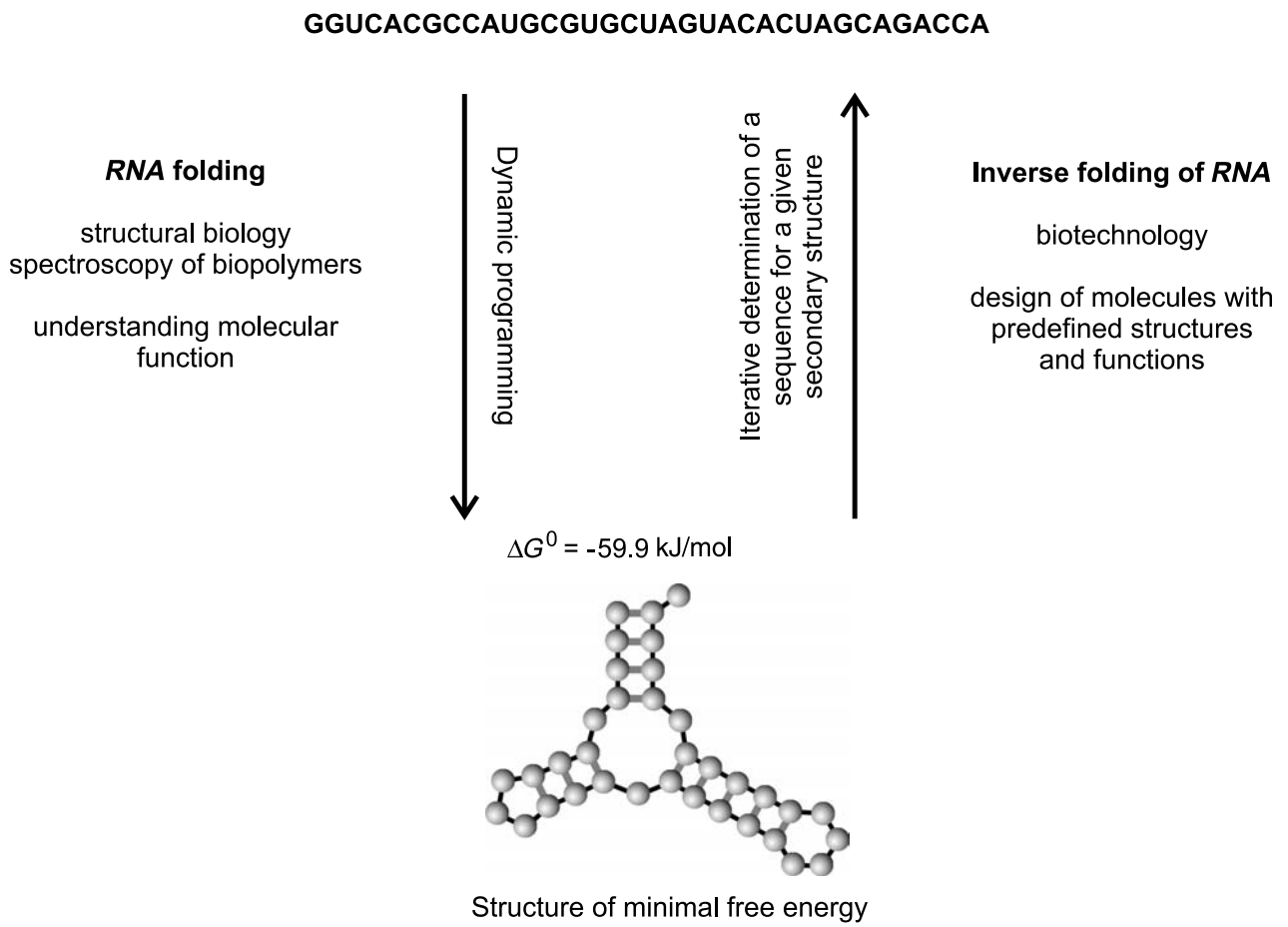

Fig. 5 Folding and inverse folding of *RNA*. *RNA* folding assigns a structure, commonly the mfe structure, to every sequence (l.h.s.). In the structure design problem (r.h.s.) a sequence is calculated that forms the given structure as its mfe structure. This inverse folding problem is solved in an iterative way [50]. In general, inverse folding is not unique in the sense that many sequences form the same mfe structure

out that a third move, the shift move, is required in which a nucleotide shifts directly between two pairing partners[8]. The algorithm applied to the calculations of trajectories simulates the *Markov* process underlying the master equation for the chemical reaction network [152, 153]. These direct simulations provide important insights into *RNA* structure formation, but can only be performed for rather small molecules because of the enormous computational efforts required in both CPU-time and memory. An alternative approach starts from the complete set of suboptimal conformations. Using the conventional *Arrhenius* formula, reaction rate parameters can be calculated for every elementary step and the folding kinetics can be calculated by solving the reaction network of conformational changes [55]. Further simplification restricts transitions to processes on the barrier tree.

The move set defines a distance $d_{1,2}^{(s)}$ between two conformations, S_1 and S_2, as the minimum number of moves that is required to convert S_1 into S_2. It is straightforward to verify that $d_{1,2}^{(s)}$ fulfils the conditions for a metric in the space of conformations. Assignment of free energies to the points representing the conformations results in a free energy landscape. Each move set defines a (multi-dimensional) folding landscape, since different move sets give rise to different neighborhood relations and therefore induce different landscapes in conformation space. The barrier tree can be understood as a ''one-dimensional'' approximation to the folding landscape (for more details on kinetic folding of *RNA* secondary

[8] The shift move can be understood as a combination of base pair opening and base pair formation in a single step.

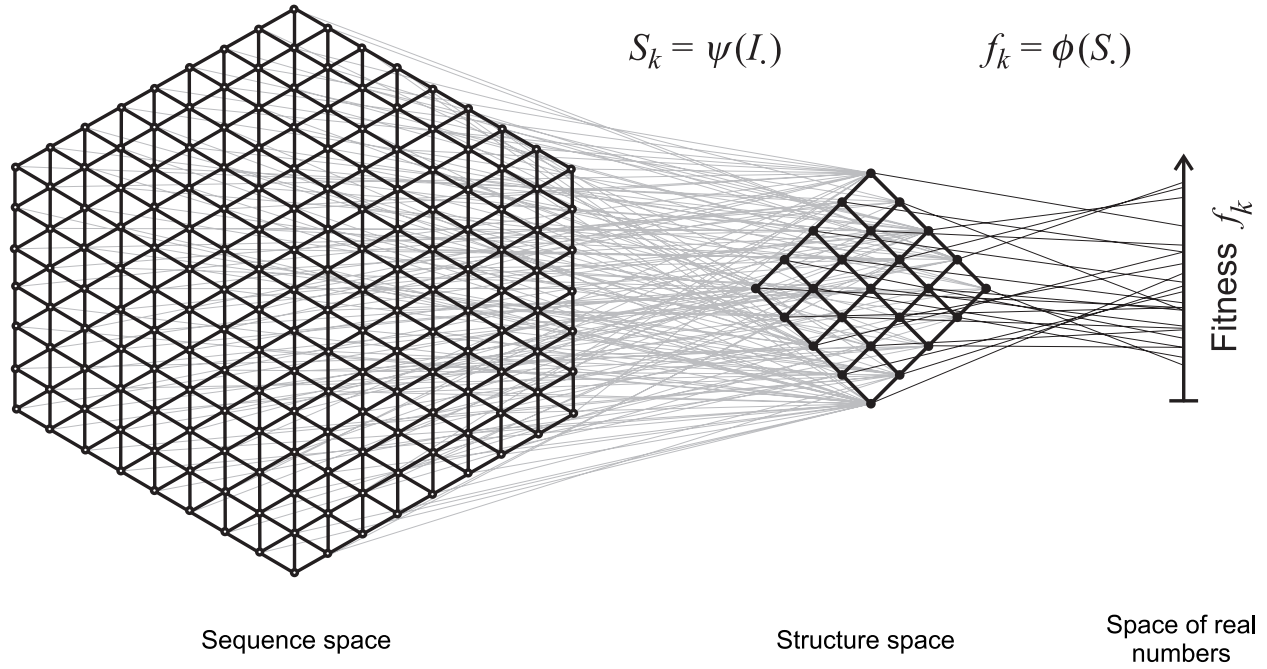

Fig. 6 Combinatory map of *RNA* structures. The relation between sequences and structures as modeled as a mapping from sequence space into structure space, $S_k = \psi(I.)$. This map is not invertible, because we are dealing almost always with many more sequences than structures. Scalar properties of molecular function can be visualized as a second mapping from structure space into the real numbers, $f_k = \phi(S.)$.

structures see the contribution by *Hofacker* and *Flamm* in this issue).

Structure design

The inversion of *RNA* secondary structure prediction is the structure design problem: the computation of sequences that form a given secondary structure as the minimum free energy structure. Inverse problems are often solved by iteration of the forward problem and such a strategy is used here as well. The inverse folding algorithm for *RNA* secondary structures starts from a (randomly) chosen sequence, which is compatible with the structure[9]. The sequence is changed by single nucleotide mutations in such a way that the distance between the current mfe structure and the predefined mfe structure decreases. Different initial sequences, in essence, give rise to different solutions as the inverse folding problem is not unique. It may also happen that the algorithm finds no solution to some initial sequences. Since there exist structures that cannot be formed by any sequence of the given length [154] inverse folding may also have no solution.

The design of multistable *RNA* molecules is more tricky. The design problem can be transformed into a combinatorial optimization problem and solved by means of a simple heuristic [155]. Molecules with two dominant conformations called *RNA* switches like the one shown in Fig. 4 are readily obtained by this procedure indicating that such conformational switches should also be accessible in evolution.

Sequence structure mapping

In general, $N_s = 4^n$ different sequences are possible for *RNA* molecules with a chain length of n nucleotides, whereas the number of secondary structures is definitely smaller than $N_{\mathrm{str}} \ll 3^n$. A calculation based on combinatorics yields the asymptotic expression for long chain lengths $N_{\mathrm{str}} \approx 1.4848 \times n^{-3/2}(1.84892)^n$ [156]. For polymers ($n \geq 50$) the numbers of sequences exceed the numbers of structures by many orders of magnitude and hence, we expect to encounter extensive neutrality with respect to minimum free energy structure formation [57]. In

[9] A sequence is called compatible with a structure when it has an admissible base combination (**A** and **U**, **C** and **G**, **G** and **C**, **G** and **U**, **U** and **A**, or **U** and **G**) at all positions where base pairs occur in the structure.

other words the cardinality of sequence space is enormous compared to that of structure space. In particular, this is even true for binary sequences[10] where we are dealing with $N_s = 2^n$ sequences of chain length n.

The sequences folding the same mfe structure S form *a neutral network* in sequence space [157]. A neutral network is a graph in sequence space with $G[S]$ being the nodes:

$$G[S] = \psi^{-1}(S) = \{I|\psi(I) = S\}, \qquad (1)$$

The edges connect all pairs of sequences belonging to this graph that are converted into each other by a single point mutation[11]. A useful quantity for the characterization of neutral networks is the degree of neutrality, $\bar{\lambda}$, which is obtained by averaging the fraction of *Hamming* distance one neighbors that form the same mfe structure, $\lambda_I = n_{\text{ntr}}^{(1)}/(n \cdot (\kappa - 1))$ with $n_{\text{ntr}}^{(1)}$ being the number of neutral one-error neighbors, over the whole network, $\mathbf{G}[S]$:

$$\bar{\lambda}[S] = \frac{1}{|\mathbf{G}(S)|} \sum_{I \in \mathbf{G}[S]} \lambda_I \qquad (2)$$

Connectedness of neutral networks is, among other properties, determined by the degree of neutrality, [157]:
With probability one a network is:

$$\begin{cases} \text{connected} & \text{if } \bar{\lambda} > \lambda_{\text{cr}} \\ \text{not connected} & \text{if } \bar{\lambda} > \lambda_{\text{cr}} \end{cases} \qquad (3)$$

where $\lambda_{\text{cr}} = 1 - \kappa^{\frac{1}{\kappa - 1}}$ where κ is the number of letters in the nucleotide alphabet, $\kappa = 4$ for the natural (**AGCU**)-alphabet and $\kappa = 2$ for binary sequences. Computations yield $\lambda_{\text{cr}} = 0.5$, 0.423, and 0.370 for the critical value in two, three, and four letter alphabets. It is remarkable that the connectivity threshold depends exclusively on the number of digits in the nucleotide alphabet. Random graph theory predicts a single largest component for non connected networks, *i.e.*, networks below threshold, that is commonly called the 'giant component'. Real neutral networks derived from *RNA* secondary structures may deviate from the prediction of random graph theory in the sense that they have two or four equally sized largest components [58, 154].

Dynamical systems of genetic regulation networks

Dynamical systems theory provides highly useful tools for the analysis of qualitative behavior of low-dimensional differential equations. Chemical reaction networks are commonly modeled by dynamical systems consisting of autonomous ordinary differential equations (ODEs) of the type

$$\frac{dx}{dt} = \dot{x} = f(x;p), \quad x \in D \subset R^n \text{ and } \quad p \in R^m, \quad (4)$$

where $x = (x_1, \ldots, x_n)$ are the concentration variables and $p = (p_1, \ldots, p_m)$ the parameters[12]. The domain of concentration variables D is always a subset of the real numbers, because concentrations by definition are nonnegative numbers. The vector f subsumes the right hand side of the kinetic equations:

$$f(x;p) = \begin{pmatrix} f_1(x;p) \\ f_2(x;p) \\ \vdots \\ f_n(x;p) \end{pmatrix}.$$

For a given set of parameters the complete set of initial conditions x_0 determines uniquely a solution curve or trajectory of the dynamical system. Uniqueness implies that trajectories never cross. The trajectories end in ω-limits, which are called attractors and may consist of single points as well as manifolds of two or more dimensions like limit cycles, chaotic attractors, or they may also diverge in an infinite domain. Time reversal, *i.e.*, replacing t by -t, causes the trajectories to converge to the α-limits. The set of all trajectories of a dynamical system is called the phase portrait and defines a flux that leads from α- to ω-limits. Nonlinear dynamical systems commonly have more than one attractor. Then, the domain D is partitioned into basins of attraction, which are separated by *separatrices* (Fig. 7). When trajectories cross the boundary of the positive orthant they can only do it the direction from outside to inside, since concentrations can never become negative.

[10] Binary sequences oligo- or polynucleotide sequences contain only two mononucleotides that can form a base pair, **C** and **G**, **A** and **U**, or **D** and **U** where **D** is 2,6-diamino-purine.
[11] These are the pairs of sequences with *Hamming* distance $d_H = 1$. The *Hamming* distance is the minimal number of mutations to convert one sequence into the other.
[12] For simplicity column and row vectors are distinguished only when it is necessary. Here $\dot{x}$ is a column vector what becomes clear from the definition of $f(x;p)$.

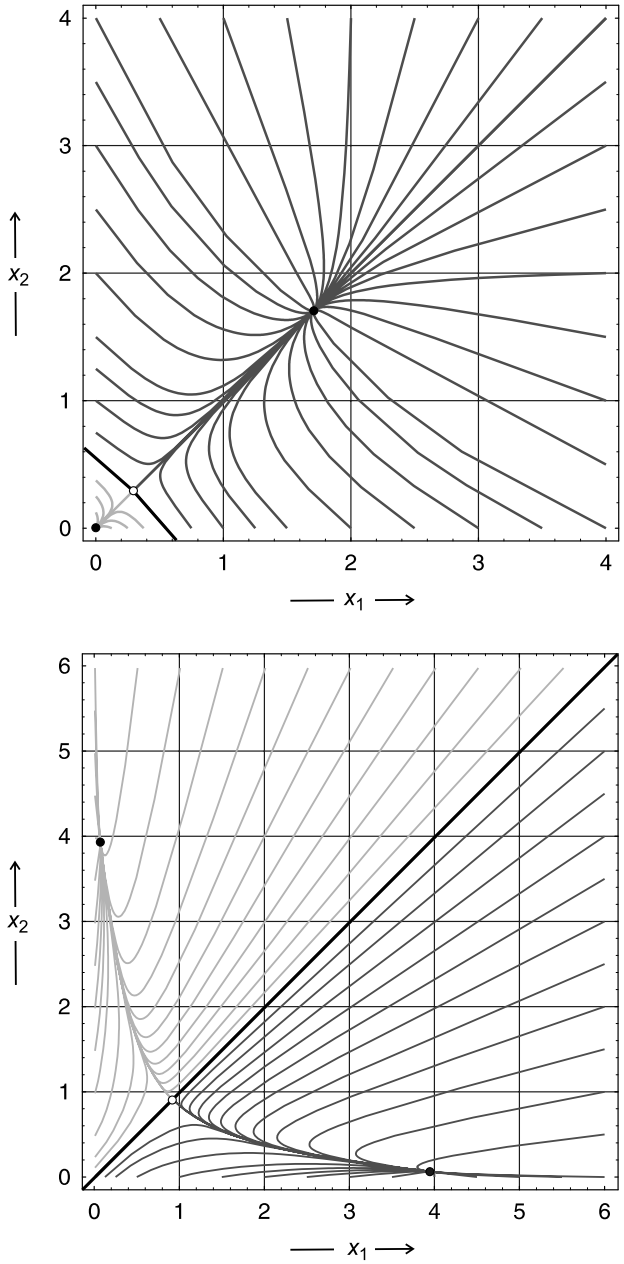

Fig. 7 Phase portraits of gene regulation. The upper part of the figure shows the case of cross-activation of two genes. Two stable points (*black circles*) are separated by a separatrix (*black curve*) that passes through an unstable point (*a saddle*, *white circle*). The stable states represent the alternatives both genes on (*dark grey trajectories*) and both genes off (*light grey trajectories*). The lower part shows cross-repression leading to the alternatives gene 1 on and gene 2 off (*dark grey*) and gene 1 off and gene 2 on (light grey). A *Hill* coefficient of $n=2$ and the values $k_1=k_2=2$, $K_1=K_2=0.5$, and $d_1=d_2=1$ (*upper plot*), and $k_1=k_2=4$, $K_1=K_2=0.25$, and $d_1=d_2=1$ (*lower plot*) for the parameters were chosen.

Gene regulation

As an example of bifurcation analysis applied to genetic regulation we present a dynamical system describing transcription and translation of two regulatory proteins, $\mathbf{P}_1$ and $\mathbf{P}_2$, controlling the activity of two genes $\mathbf{G}_1$ and $\mathbf{G}_2$ through cross interaction, *e.g.*, $\mathbf{G}_1$ is controlled by $\mathbf{P}_2$ and $\mathbf{G}_2$ by $\mathbf{P}_1$, respectively. The transcription products of the genes are assumed to be two messenger *RNA*s, $\mathbf{Q}_1$ and $\mathbf{Q}_2$, which encode the amino acid sequences of the two proteins. Neglecting all intermediates the over-all reaction mechanism for transcription, translation, and degradation has the simple form

$$\begin{aligned}
&\mathbf{G}_i + \mathbf{P}_j \overset{k_i}{\rightleftharpoons} \mathbf{G}_i \cdot \mathbf{P}_j \xrightarrow{k_i^Q} \mathbf{G}_i \cdot \mathbf{P}_j + Q_i && \text{activation}\\
&\mathbf{G}_i + \mathbf{Q}_i \xleftarrow{k_i^Q} \mathbf{G}_i + \mathbf{P}_j \overset{k_i}{\rightleftharpoons} \mathbf{G}_i \cdot \mathbf{P}_j && \text{repression}\\
&\mathbf{Q}_i \xrightarrow{k_i^Q} \mathbf{Q}_i + \mathbf{P}_i && \text{translation}\\
&\mathbf{Q}_i \xrightarrow{d_i^Q} \emptyset && \text{degradation}\\
&\mathbf{P}_i \xrightarrow{d_i^P} \emptyset && \text{degradation}
\end{aligned} \tag{5}$$

with $i=1,\ 2$ and $j=2,\ 1$. Activation and repression refer here to the action of the regulatory protein on the gene: binding of the activator is required for gene activity whereas repressor binding prevents transcription. Concentration variables are denoted by lower case letters, $[\mathbf{Q}_i]=q_i$ and $[\mathbf{P}_i]=p_i$ $(i=1,\ 2)$.

The equilibrium parameters, K_1 and K_2, are given as dissociation constants, and therefore lower values of K refer to stronger binding. The kinetic rate parameters are denoted by k and d, the superscripts, 'Q' and 'P' refer to m*RNA*s and proteins, respectively. Then the kinetic equations including the degradation terms for m*RNA*s and proteins are of the form:

$$\begin{aligned}
\frac{dq_i}{dt} &= k_i^Q g_i^0 F_i(p_j) - d_i^Q q_i, \quad i=1,2 \ \text{ and } \ j=2,1\\
\frac{dp_i}{dt} &= k_i^P q_i - d_i^P p_i, \quad i=1,2.
\end{aligned} \tag{6}$$

Activation and repression are commonly modeled by means of simple binding functions for complex formation:

$$\begin{aligned}
F^{\text{act}}(p) &= \frac{p^n}{K+p^n} && \text{activation}\\
F^{\text{rep}}(p) &= \frac{K}{K+p^n} && \text{repression},
\end{aligned} \tag{7}$$

where n is the *Hill* coefficient that is related to cooperative binding of multimeric proteins [158] and proteins as *DNA* replication is not considered here, and the total concentration of the genes, $[\mathbf{G}_i] = g_i^0$; $(i = 1, 2)$, is a constant and can be subsumed in the rate parameter $(k_i^Q \doteq k_i^Q g_i^0)$.

For the purpose of illustration we can make another simplification that has only little consequences for the dynamical behavior around stationary points: protein concentrations are assumed to be proportional to the m*RNA* concentrations ($p_1 = \kappa_1 q_1$ with $\kappa_i = k_i^P / d_i^P$), which becomes exact at the stationary states. The kinetic equations for the concentrations of m*RNA*s are formally unchanged when the dissociation constants are properly scaled: $K_\mathrm{i} \Rightarrow K_\mathrm{i}/\kappa_i^n$ $(i = 1, 2)$. Concentration space is then only two dimensional and trajectories can be visualized in a plane[13]. The phase portraits in Fig. 7 were calculated for cross-regulation of the two genes leading for activation to two states with both genes on or both genes off, and to a toggle switch, gene 1 on, gene 2 off or gene 1 off, gene 2 on, for the repression case, respectively. The approximation describes well only the situation near stationary states and thus the phase portraits contain only point attractors. It fails, however, for oscillatory systems, which have limit cycle attractors at substantial distance from a stationary point (see below).

The phase portrait of a dynamical system depends on the system parameters, which are the equilibrium constant of regulatory complexes and the reaction rate parameters in the example reported here. Commonly phase portraits stay qualitatively the same for large variations in the parameters[14] but then change abruptly through bifurcations at certain parameter values. In order to illustrate bifurcation behavior it is necessary to identify one parameter or one characteristic parameter combination for the variation, for example the transcription rate parameter k_i, the complex dissociation constant K_i, or both. Figure 8 presents an example of a pitchfork bifurcation at which the qualitative behavior of the system changes: one stable stationary point lying on the symmetry axis $x_1 = x_2$ is replaced by an unstable stationary point on the symmetry axis and two symmetrically lying stable points, $\bar{x}_1 > \bar{x}_2$ and $\bar{x}_2 > \bar{x}_1$. The toggle switch discussed in the previous paragraph thus requires a situation beyond the bifurcation point. As seen from the numerical values used for the parameters, strong binding and fast transcription favor the genetic switch.

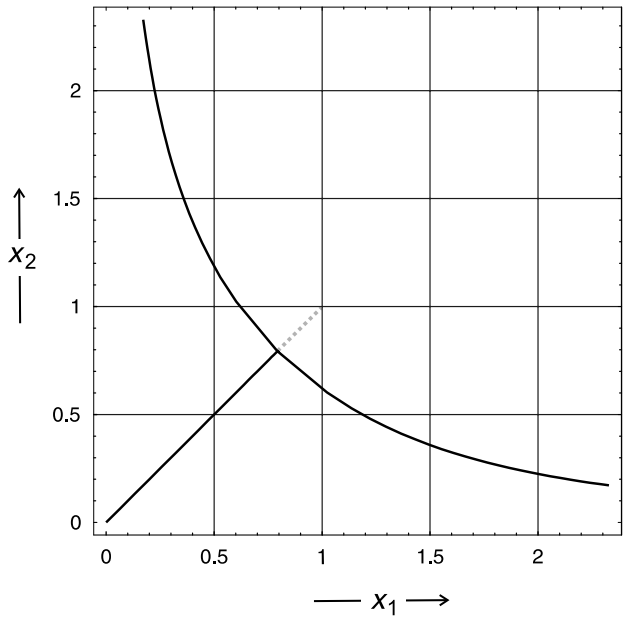

Fig. 8 A pitchfork bifurcation in gene regulation. The figure shows the dependence of stationary points in the repression-repression case with *Hill* coefficient $n = 2$. Variation of the parameters was introduced by means of an auxiliary variable s: $k_1 = k_2 = 1 \cdot s$ and $K_1 = K_2 = 1/s$ $(d_1 = d_2 = 1)$. The pitchfork bifurcation is observed at: $s_\mathrm{crit} = 1.58746$. Below the critical point, $s < s_\mathrm{crit}$, one stable stationary point with $\bar{x}_1 = \bar{x}_2$ is observed, whereas the stationary point is unstable and two other stable stationary points exist above the critical value, $s > s_\mathrm{crit}$, as shown in the phase portrait in Fig. 7

Elimination of variables by means of a stationarity assumption may lead to completely wrong model behavior in regions far off the stationary points. For example, the combination of activation and repression, $F_i^\mathrm{act}(p_j)$ and $F_j^\mathrm{rep}(p_i)$, gives rise to a *Hopf* bifurcation and undamped oscillations at sufficiently strong binding and sufficiently large kinetic parameters for *Hill* coefficients $n \geq 3$ [159]. In the simplified system with two variables (x_1, x_2) no undamped oscillations occur, which can be proven by straightforward calculations.

[13] In order to indicate the assumption of proportionality the variables are denoted by x_1 and x_2 and the superscripts 'Q' and 'P' are dropped on the rate constants.

[14] In precise mathematical terms qualitative identity of two phase portraits means that they are related by a *homeomorphism* or continuous transformation implying an equivalence relation and one-two-one correspondence between points in the two figures that is continuous in both directions. In particular, this is expressed in the same topological relations between attractors and separatrices.

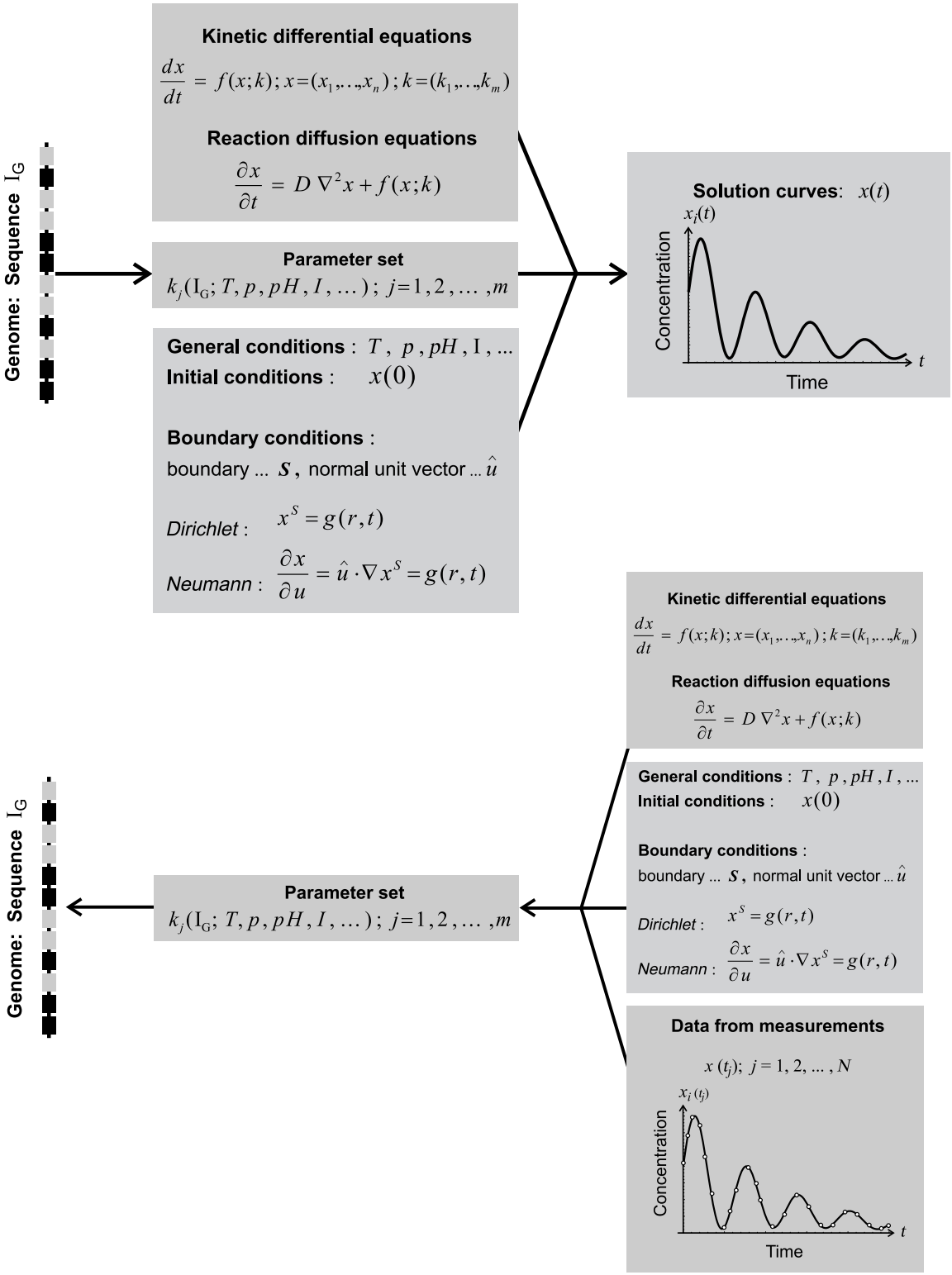

Fig. 9 Forward and inverse methods in biochemical kinetics. The two diagrams sketch typical forward (*upper part*) and inverse (*lower part*) problems in systems biology. Dynamics is modeled by means of ordinary or partial differential equations. In the forward problem solution curves are computed from known model equations, parameters, and conditions. The inverse problem determines parameters from model equations, conditions, and measurements. In systems biology parameters are derived ultimately from genomics and proteomics data, or they allow for making inferences on genetic and metabolic systems from known parameter values

Inverse methods

Conventional techniques of modeling chemical reaction networks by means of differential equations are based on the forward approach of reaction kinetics (Fig. 9): kinetic equations, general, boundary, and initial conditions, as well as the parameters are assumed to be known, solution curves are computed and compared to experimental data. Unknown parameters are commonly determined by fitting to data that were measured under suitable conditions. Inverse problems became first popular in scattering theory: the angular intensity distribution of scattered radiation is recorded after a scattering event and the scattering object has to be reconstructed. Highly elaborate and fully automated methods for this reconstruction are available in computer tomography (CT) and magnetic resonance imaging (MRI). The inverse problem in reaction kinetics is concerned with the direct determination of parameters from data. In particular, a set of experimental data is given as input and the set of parameters is determined from known mechanism and conditions (Fig. 9). Apart from exceptional pathological cases forward problems are well-posed in the sense of *Jacques Salomon Hadamard*[15]. Inverse problems are almost always ill-posed. The case of inverse folding of *RNA* has been mentioned already: when properly formulated the forward problem has a unique solution, which is not the case for the inverse problem [57]. The same is true, in essence, for parameter identification of kinetic differential equations. Ill-posed problems require unconventional or special techniques for finding approximative solutions, regularization with several variants is most frequently used [160, 161]. For nonlinear systems the inverse problem is solved by iterations. A general overview on solution methods for inverse problems and applications to some selected problems are found in the two collective volumes by *Engl et al.* [162] and *Colton et al.* [163].

Experimental data always contain a certain amount of noise that gives rise to uncertainty in the determined parameter values. Often the data are not sufficient for providing reliable information on all parameters. Sensitivity analysis and application of sparsity constraints[16] are suitable tools for the identification of relevant and faithfully predicable parameters.

Reverse engineering of bifurcation behavior – also characterized as a level two inverse method applied to dynamical systems – aims at the design of a predefined bifurcation pattern. An algorithmic procedure for inverse bifurcation analysis has been conceived and automated for iterative computation of approximative solutions [164]. We sketch the basic idea of the approach: the m-dimensional parameter space of Eq. (4), $P \subset \mathrm{R}^m$, is partitioned into input and system parameters: $p = (p_i, p_s) \in P_i \times P_s$. The bifurcation manifold Σ consists of sets in parameter space P for which structural stability breaks down [165]. For a value of the system parameter p_s we define $\Sigma\,(p_s) = \Sigma \cap \{p_s\}$ being the intersection of Σ with the plane defined by p_s (Fig. 10). The forward problem consists in finding the orthogonal projection of some point p in parameter space onto the manifold $\Sigma(p_s)$. In other word the forward operator is a mapping $F\colon P \rightarrow P$ in parameter space that can be assumed to be well-posed:

$$F(p) \equiv (F(p)_i, F(p)_s) = (\mathcal{P}_{\perp[\Sigma(p_s)]} p_i, p_s) \qquad (8)$$

Here $\mathrm{P}_{\perp[\Gamma]}$ is an orthogonal projection operator onto the manifold Γ. Figure 10 shows an iterative procedure calculating $F(p)$ in case of a nonlinear system.

The inverse bifurcation problem consists in a variation of p_s with the goal to bring the point p_i as close as possible to the bifurcation manifold. In mathematical terms it is formulated by

$$\min_{p_s} J(p) = \|F(p)_i - p_i\| \quad \text{subject to}$$
$$p_{\mathrm{low}} \leq p \leq p_{\mathrm{upp}} \quad \text{and} \quad 0 \leq \gamma(F(p)_i), \qquad (9)$$

where $\|\cdot\|$ is the l_2 norm and $\gamma : P_i \rightarrow R^k$ represents some k-dimensional nonlinear constraints. The region to be searched in parameter space is commonly bounded by physical or chemical restrictions resulting in lower and upper bounds in the parameters, which are appropriately introduced into the algorithm.

The method described here has been used to perform reverse engineering for a number of relevant biological problems [160]. Examples are the optimization of circadian rhythms with respect to insensitivity to temperature, the conditions under which the cell cycle in yeast can be locked in the S-phase, the choice of parameters that allows for oscillations in gene regulatory systems, and maximization of oscillatory regimes in parameter space.

[15] For a well-posed problem (i) a solution exists, (ii) the solution is unique, and (iii) the solution depends continuously on the data in some reasonable topology.

[16] Sparsity means that many parameters take on very small values. Application of a sparsity constraint implies that all parameter up to a certain threshold value are set zero.

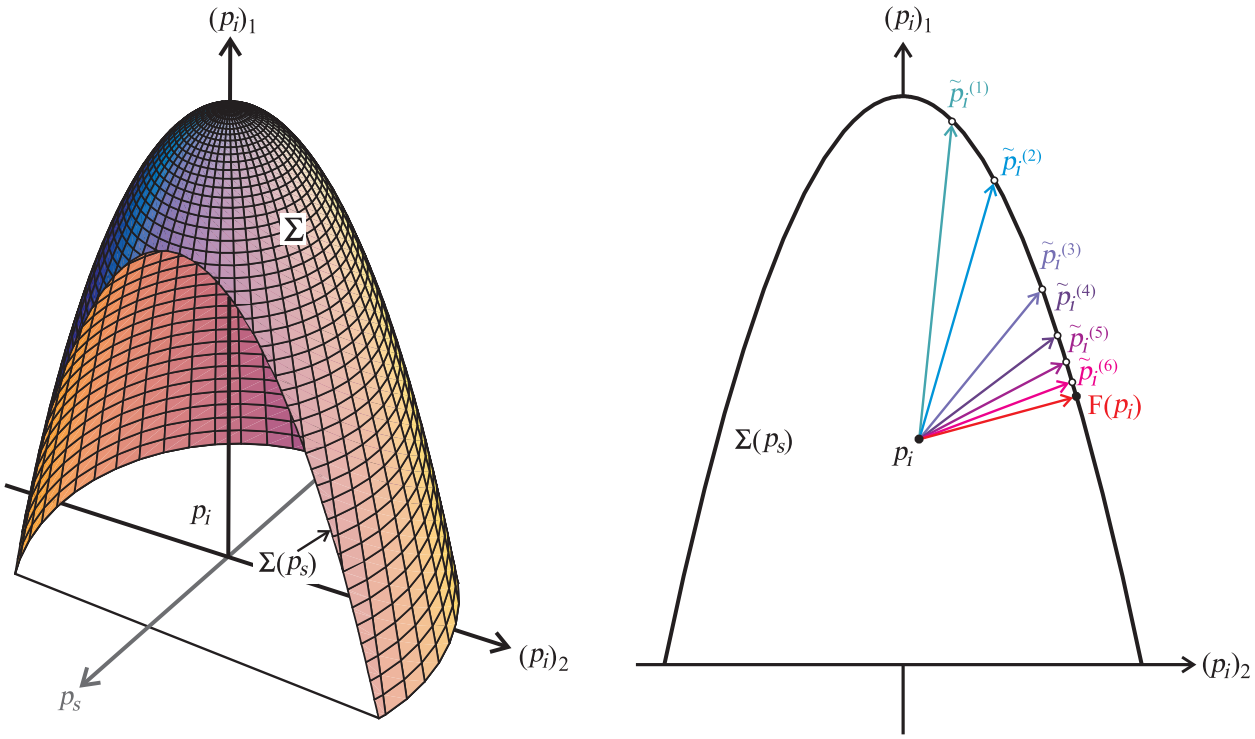

Fig. 10 Inverse bifurcation analysis. Parameter space is partitioned into a space of input parameters, $p_i = ((p_i)_1, (p_i)_2, \dots)$ and a space of system parameters p_s. The dynamical system contains a bifurcation manifold Σ that has the intersection manifold $\Sigma(p_s) = \Sigma \cap \{p_s\}$ with the space of system parameters (l.h.s. of the figure). In the forward problem we search for the point $F(p_i)$ that is closest to some point p_i in the input parameter space. In case of nonlinear dynamical systems this point is computed through iterations on the manifold $\sum(p_s) : p_i \to \tilde{p}_i^{(1)} \to \tilde{p}_i^{(2)} \to \tilde{p}_i^{(3)} \to \dots$ (r.h.s. of the figure).

Concluding remarks

Biochemistry, molecular biology, and genome research are currently reaching a point where rigorous mathematical methods and efficient computational techniques can be applied. Thereby biological modeling can be placed upon a firm molecular basis. Still many problems have to be solved and open questions remain for principal issues. Examples are the handling of low particle numbers and fluctuations, the description of spatial heterogeneity or the analysis of processes involving multi-component supramolecular complexes to mention just the most obvious problems that call for novel approaches. Indeed, the mechanisms, by which natural nanodevices or molecular machines perform the most complex cellular processes, are largely unknown. Nevertheless, present day biology has become firmly rooted in chemistry and physics without loosing its specific approach towards understanding nature and the enormously rich wealth of observations and data provides for the first time a fundament upon which a theoretical biology of the future can be placed. In view of the breathtaking progress of knowledge and data accumulation in current biology the need for a comprehensive theory of cellular life based on structural biology and chemical kinetics becomes more and more urgent every day.

Acknowledgements

The work reported here was supported financially by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung, (Project No. 14898-MAT), by the European Commission (Project No. PL970189), by the Wiener Wissenschafts-, Forschungs- und Technologiefonds (Project No. MA05), and by the Santa Fe Institute.

References

1. Hartl DL, Clark AG (1997) Principles of Population Genetics, 3rd edn. Sinauer Associates, Sunderland, MA
2. Kimura M (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK
3. Michaelies L, Menten ML (1913) Biochemische Zeitschrift 49:333
4. Friess SL, Lewis ES, Weissberger A (1963) Investigation of Rates and Mechanisms of Reactions, Vol. VIII – Part II of Technique of Organic Chemistry, 2nd edn. Interscience, New York

5. Schwarz G (1968) Rev Mod Phys 40:206
6. Flory PJ (1969) Statistical Mechanics of Chain Molecules. Interscience Publ., New York
7. Zimm BH, Bragg JK (1959) J Chem Phys 31:526
8. Eigen M, Maeyer L de (1963) Relaxation Kinetics. In: Friess SL, Lewis ES, Weissberger A (eds) Technique of Organic Chemistry, Vol. VIII/2, Chapter 18, 2nd edn. Interscience Publishers, New York, p 895
9. Gutfreund H (1971) Annu Rev Biochem 40:315
10. Ising E (1925) Z Phys 31:253
11. Schwarz G (1968) Biopolymers 6:873
12. Schwarz G (1965) J Mol Biol 11:64
13. Go M, Go N, Sheraga HA (1968) I Formulation Proc Natl Acad Sci USA 59:1030
14. Go N, Go M, Sheraga HA (1970) J Chem Phys 52:2060
15. Mitsutake A, Okamoto Y (2000) J Chem Phys 112:10638
16. Chakrabartty A, Kortemme T, Baldwin RL (1994) Protein Sci 3:843
17. Klipp E, Herwig R, Kowald A, Wieling C, Lehrach H (2005) Systems Biology in Practice. Concepts, Implementation, and Application. Wiley-VCh, Weinheim, DE
18. Judson HF (1979) The Eighth Day of Creation. The Makers of the Revolution in Biology. Jonathan Cape, London
19. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) Science 289:905
20. Moore PB, Steitz TA (2003) Annu Rev Biochem 72:813
21. McMillan PF, Clary DC (2005) Phil Trans Roy Soc A 363:311
22. Rappe AK, Casewit CJ (1997) Molecular Mechanics across Chemistry. University Sciecne Books, Sausalito, CA
23. Leach AR (2001) Molecular Modelling. Principles and Applications, 2nd edn. Prentice Hall, Harlow, GB
24. Sippl MJ (1990) J Mol Biol 213:859
25. Sippl MJ (1990) J Computer-Aided Mol Design 213:859
26. Poole AM, Ranganathan R (2006) Curr Op Struct Biol 18:508
27. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Annu Rev Biophys Biomol Struct 29:291
28. Koppensteiner WA, Lackner P, Wiederstein M, Sippl MJ (2000) J Mol Biol 296:1139
29. Zhang Y, Skolnick J (2005) Proc Natl Acad Sci USA 102:1029
30. Lattman EE (2005) Proteins 61(Suppl 7):1
31. Kryshtafovych A, Venclovas C, Fidelis K, Moult J (2005) Proteins 61:225
32. Pierce NA, Winfree E (2002) Protein Engi 15:779
33. Dahiyat BI, Mayo SL (1997) Natl Acad Sci USA 94:10172
34. Street AG, Mayo SL (1999) Structure 7:R105
35. Voigt CA, Gordon DB, Mayo SL (2000) J Mol Biol 299:789
36. DeGrado WF (2001) Chem Rev 101:3025
37. Butterfoss GL, Kuhlman B (2006) Annu Rev Biophys Biomol Struct 35:49
38. Lippow SM, Tidor B (2007) Curr Op Biotech 18:305
39. Drew HR, Wing RM, Takano T, Broka C, Tanaka S, Itakura K, Dickerson RE (1981) Proc Natl Acad Sci USA 78:2179
40. Dickerson RE, Drew HR (1981) J Mol Biol 149:761
41. Drew HR, Dickerson RE (1981) J Mol Biol 151:535
42. Neidle S (1998) Nature Struct Biol 5:754
43. Packer MJ, Dauncey MP, Hunter CA (2000) J Mol Biol 295:85
44. Gardiner EJ, Hunter CA, Packer MJ, Palmer DS, Willett P (2003) J Mol Biol 332:1025
45. Dickerson RE, Ng HL (2001) Proc Natl Acad Sci USA 98:6986
46. Vargason JM, Henderson K, Ho PS (2001) Proc Natl Acad Sci USA 98:6986
47. Tolstorukov MY, Ivanov VI, Malenkov GG, Jernigan RL, Zhurkin VB (2001) Biophys J 81:3409
48. Ng HL, Dickerson RE (2002) Nucleic Acids Res 30:4061
49. Zuker M, Stiegler P (1981) Nucleic Acids Res 9:133
50. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LB, Tacker M, Schuster P (1994) Mh Chemie 125:167
51. Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Biopolymers 49:145
52. Zuker M (1989) Science 244:48
53. McCaskill JS (1990) Biopolymers 29:1105
54. Flamm C, Fontana W, Hofacker IL, Schuster P (1999) RNA 6:325
55. Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF (2004) J Phys A Math Gen 37:4731
56. Andronescu M, Fejes AP, Hutter F, Hoos HH, Condon A (2004) J Mol Biol 336:607
57. Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) Proc Roy Soc London B 255:279
58. Schuster P (2006) Rep Prog Phys 69:1419
59. Olson WK, Bansal M, Burley SK, Dickerson RE, Gerstein M, Harvey SC, Heinemann U, Lu XJ, Neidle S, Shakked Z, Sklenar H, Suzuki M, Tung CS, Westhof E, Wolberger C, Berman HM (2001) J Mol Biol 313:229
60. Leontis NB, Westhof E (2001) RNA 7:499
61. Lascoute A, Leontis NB, Massire C, Westhof E (2005) Nucleic Acids Res 33:2395
62. Leontis NB, Lescoute A, Westhof E (2006) Curr Op Struct Biol 16:279
63. Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, Engelke DR, Harvey SC, Holbrook SR, Jossinet F, Lewis SE, Major F, Mathews DH, Richardson JS, Williamson JR, Westhof E (2006) RNA 12:553
64. Brakmann S, Johnsson K (2002) Directed Molecular Evolution of Proteins or How to Improve Enzymes for Biocatalysis. Wiley-VCH, Weinheim, DE
65. Klussmann S (ed) (2006) The Aptamer Handbook. Functional Oligonucleotides and Their Applications, Wiley-VCh Verlag, Weinheim, DE
66. Ellington AD, Szostak JW (1990) Nature 346:818
67. Tuerk C, Gold L (1990) Science 249:505
68. Eigen M (1971) Naturwissenschaften 58:465
69. Eigen M, Schuster P (1977) Naturwissenschaften 64:541

70. Eigen M, McCaskill J, Schuster P (1989) Adv Chem Phys 75:149
71. Fontana W, Schuster P (1987) Biophys Chem 26:123
72. Fontana W, Schuster P (1998) Science 280:1451
73. Fontana W, Schuster P (1998) J Theor Biol 194:491
74. Maxam A, Gilbert W (1977) Proc Natl Acad Sci USA 74:560
75. Sanger F, Nicklen S, Coulson AR (1977) Proc Natl Acad Sci USA 74:5463
76. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell JR, Heiner C, Kant SBH, Hood LE (1986) Nature 321:674
77. Weber JL, Myers EW (1997) Genome Res 7:401
78. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei YD, Li ZY, Li JY, Liang Y, Lin XY, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue BX, Sun JT, Wang ZY, Wang AH, Wang X, Wang J, Wei MH, Wides R, Xiao CL, Yan CH, Yao A, Ye J, Zhan M, Zhang WQ, Zhang HY, Zhao Q, Zheng LS, Zhong F, Zhong WY, Zhu SPC, Zhao SY, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An HJ, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi HY, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays AD, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu XJ, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen MY, Wu D, Wu M, Xia A, Zandieh A, Zhu XH (2001) Science 291:1304
79. Eigen M, Rigler R (1994) Proc Natl Acad Sci 91:5740
80. Tamarat PH, Maali A, Lounis B, Orrit M (2000) J Phys Chem A 104:1
81. Rigler R, Orrit M, Basche T (eds) (2001) Single Molecule Spectroscopy. Springer, Berlin
82. Bohmer M, Enderlein J (2003) Chem Phys Chem 4:792
83. Barkai E, Jung YJ, Silbey R (2004) Annu Rev Phys Chem 55:457
84. Rigler R, Seela F (2001) J Biotechnology 86:161
85. Vercoutere WA, Winters-Hilt S, Olsen HE, Deamer DW, Haussler D, Akeson M (2001) Nature Biotech 19:248
86. Vercoutere WA, Winters-Hilt S, DeGuzman VS, Deamer DW, Ridino SE, Rodgers JT, Olsen HE, Aarziali A, Akeson M (2003) Nucleic Acids Res 31:1311
87. Braslavsky I, Herbert B, Kartalov E, Quake SR (2003) Proc Natl Acad Sci USA 100:3960
88. Astier Y, Braha O, Bayley H (2006) J Am Chem Soc 128:1705
89. ENCODE Project Consortium (2007) Nature 447:799
90. Greally JM (2007) Nature 447:782
91. Page RDM, Holmes EC (1998) Molecular Evolution. A Phylogenetic Approach. Blackwell Science, Oxford, UK
92. Neddleman SB, Wunsch CD (1970) J Mol Biol 48:443
93. Smith TF, Waterman MS (1981) Adv Appl Math 2:482
94. Mount DW (2001) Bioinformatics. Sequence and Genome Analysis, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
95. Lawrence JG, Ochman H (1998) Proc Natl Acad Sci USA 95:9413
96. Gogarten JP, Doolittle WF, Lawrence JG (2002) Mol Biol Evol 19:2226
97. Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Trends Genet 18:472
98. Doolittle WF (1999) Science 284:2124
99. Huynen MA, Snel B, Bork P, Stiller JW, Hall BD, Gupta RS, Soltys BJ (1999) Science 286:1443
100. Martin W (1999) BioEssays 21:99
101. Rivera MC, Lake JA (2004) Nature 431:152
102. Bapteste E, Susko E, Leigh J, MacLoed D, Charlebois RL, Doolittle WF (2005) BMC Evolut Biol 5:33
103. Philippe H, Douady CJ (2003) Curr Op Microbiol 6:498
104. Doolittle RF (2005) Curr Op Struct Biol 15:248
105. Grunewald S, Forstlund K, Dress A, Moulton V (2007) Mol Biol Evol 24:532
106. Dopazo J, Dress A, von Haeseler A (1993) Proc Natl Acad Sci USA 90:10320
107. Eigen M, Winkler-Oswatitsch R, Dress A (1988) Proc Natl Acad Sci USA 85:5913
108. Weissmann C (1974) FEBS Lett 40:S10

109. Sankoff D (1985) SIAM J Appl Math 45:810
110. Gorodkin J, Heyer LJ, Stormo GD (1997) Nucl Acids Res 25:3724
111. Mathews DH, Turner DH (2002) J Mol Biol 317:191
112. Mathews DH (2005) Bioinformatics 21:2246
113. Harmanci A, Sharma G, Mathews DH (2007) BMC Bioinformatics 8:130
114. Holmes I (2005) BMC Bioinformatics 6:73
115. Dowell RD, Eddy SR (2006) BMC Bioinformatics 7:400
116. Kiryu H, Tabei Y, Kin T, Asai K (2007) Bioinformatics 23:1588
117. Hull Havgaard JH, Lyngso R, Stormo GD, Gorodkin J (2005) Bioinformatics 21:1815
118. Torarinsson E, Havgaard JH, Gorodkin J (2007) Bioinformatics 23:926
119. Hofacker IL, Bernhart SHF, Stadler PF (2004) Bioinformatics 20:2222
120. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R (2007) PLoS Comp Biol 3:e65
121. Yao Z, Weinberg Z, Ruzzo WL (2006) Bioinformatics 22:445
122. Reeder J, Giegerich R (2005) Bioinformatics 21:3516
123. Tyers M, Mann M (2003) Nature 422:193
124. Zhu H, Bilgin M, Snyder M (2003) Annu Rev Biochem 72:783
125. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkaro D, Shalon D, Botstein D, Brown PO (1999) Science 283:83
126. Aebersold R, Mann M (2003) Nature 422:198
127. Young K (1998) Biol Reprod 58:302
128. Joung J, Ramm E, Pabo C (2000) Proc Natl Acad Sci USA 97:12271
129. De Jong H (2002) J Comput Biol 9:67
130. Hynne F, Dano S, Sorensen PG (2001) Biophys Chem 94:121
131. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin LI, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) Bioinformatics 19:524
132. Hindmarsh AC, Cohen SD (1996) Comput Phys 10:138
133. De Jong H, Gouze JL, Hernandez C, Page M, Sari T, Geiselmann J (2004) Bull Math Biol 66:301
134. Schilling CH, Palsson BO (1998) Proc Natl Acad Sci USA 95:4193
135. Edwards JS, Palsson BO (2000) Proc Nat Acad Sci USA 97:5528
136. Edwards JS, Ibarra RU, Palsson BO (2001) Nat Biotechn 19:125
137. Ramakrishna R, Edwards JS, McCulloch A, Palsson BO (2001) Am J Physiol 280:R695
138. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Nature 429:92
139. Palsson BO (2006) Systems Biology. Properties of Reconstructed Networks. Cambridge University Press, New York
140. Almaas E, Kovacs Vlcsek BT, Oltvai ZN, Barabasi AL (2004) Nature 427:839
141. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Science 297:1551
142. Albert R, Barabasi AL (2002) Rev Mod Phys 74:47
143. Guimera R, Amaral LAN (2005) Nature 433:895
144. Rivas E, Eddy SR (1999) J Mol Biol 285:2053
145. Waterman MS (1978) Secondary structure of single-stranded nucleic acids. In Studies on Foundations and Combinatorics. Advances in Mathematics. Supplementary studies, Vol. 1. Academic Press, New York, p 167
146. Waterman MS, Smith TF (1978) Math Biosci 42:257
147. Mathews DH, Sabina J, Zuker M, Turner DE (1999) J Mol Biol 288:911
148. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH (2004) Proc Natl Acad Sci USA 101:7287
149. SantaLucia J Jr, Allawi HL, Seneviratne PA (1996) Biochemistry 35:3555
150. SantaLucia J Jr (1998) Proc Natl Acad Sci USA 95:1460
151. Morgan SR, Higgs PG (1996) J Chem Phys 105:7152
152. Gillespie DT (1976) J Comp Phys 22:403
153. Gillespie DT (2007) Annu Rev Phys Chem 58:35
154. Schuster P (2003) Molecular insight into the evolution of phenotypes. In: James P Crutchfield, Peter Schuster (eds) Evolutionary Dynamics – Exploring the Interplay of Accident, Selection, Neutrality, and Function. Oxford University Press, New York, 163 p
155. Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M (2001) RNA 7:254
156. Hofacker IL, Schuster P, Stadler PF (1998) Disc Appl Math 89:177
157. Reidys C, Stadler PF, Schuster P (1997). Bull Math Biol 59:339
158. Hill AV (1910) J Physiology 40 [Section 11.2.1]:iv
159. Widder S, Schicho J, Schuster P (2007) J Theor Biol 241:395
160. Engl HW, Hanke M, Neubauer A (1996) Regularization of Inverse Problems – Mathematics and its Applications. Springer-Verlag, Berlin
161. Neumaier A (1998) SIAM Rev 40:636
162. Engl HW, Louis AK, Rundell W (eds) (1997) Inverse Problems in Medical Imaging and Nondestructive Testing, Springer-Verlag, Wien
163. Colton D, Engl HW, Louis AK, McLaughlin JR, Rundell W (eds) (2000) Surveys on Solution Methods for Inverse Problems. Springer-Verlag, Wien
164. Lu J, Engl HW, Schuster P (2006) AMB Algorithms Mol Biol 1:11
165. Kuznetsov YA (2004) Elements of Applied Bifurcation Theory. Springer-Verlag, New York