CONTRIBUTIONS to SCIENCE, **6** (2): 133–142 (2010)
Institut d'Estudis Catalans, Barcelona
DOI: 10.2436/20.7010.01.91 ISSN: 1575-6343
www.cat-science.cat

research reviews

The impact of theoretical chemistry on biology

Modesto Orozco*[1,2,3], F. Javier Luque[4]

1. Department of Biochemistry and Molecular Biology, Faculty of Chemistry, University of Barcelona
2. Joint Institute of Biomedical Research - Barcelona Supercomputing Center (IRB-BSC), Research Program on Computational Biology, Barcelona Science Park, Barcelona
3. National Institute of Bioinformatics, Barcelona Science Park, Barcelona
4. Departament of Physical Chemistry and Institute of Biomedicine (IBUB), Faculty of Pharmacy, University of Barcelona

Resum. Els avenços en les bases dels mètodes teòrics i l'espectacular desenvolupament de la potència de càlcul han fet possible progressar enormement en el somni dels fundadors de la química, és a dir, ser capaços d'estudiar amb mètodes computacionals el conjunt de processos químics. Actualment, la química teòrica està completant el darrer avenç: intentar esdevenir l'eina més recent per a comprendre la naturalesa química dels éssers vius. Aquesta revisió pretén mostrar com els mètodes de la química teòrica, originalment desenvolupats per a examinar molècules petites en fase gas, han evolucionat per a assolir la complexa descripció de sistemes biològics.

Paraules clau: química teòrica · química computacional · biologia computacional · bioinformàtica

Summary. Recent advances in theoretical formalism together with the dramatic development of computer infrastructure have allowed enormous progress in achieving the dream of the founders of chemistry: to submit the majority of chemical phenomena to calculation. Currently, theoretical chemistry is working towards reaching the next step: to become the ultimate tool to understand the chemical nature of living organisms. This review summarizes how techniques originally developed to represent small molecules in the gas phase have evolved such that they are able to describe the complex behaviors of biological systems.

Keywords: theoretical chemistry · computational chemistry · computational biology · bioinformatics

Introduction

Since its scientific formulation in the 18th century, chemistry has aimed at achieving the rationalization of chemical phenomena, thus abandoning the purely empirical nature of alchemy. It was in 1888 that Gay-Lussac, one of the founders of modern chemistry, stated: "We are perhaps not far removed from the time when we shall be able to submit the bulk of chemical phenomena to calculation." This was clearly overly optimistic, but illustrates the goal of chemists to not only describe but also to deeply understand chemical systems. In fact, the fathers of chemistry, including Avogadro, Boyle, Dalton, Gay-Lussac, Lavoisier, and Volta, did not enter the history books by reporting details of the behavior of chemical systems, but because of their contributions to the development of a theoretical framework able to rationalize the behavior of chemical phenomena.

The greatest theoretical challenge in chemistry is understanding the behavior of systems at the level of the atom by using the basic rules of physics. The development of quantum chemistry during the last years of the 19th century and the first half of the 20th century provided the required theoretical framework. Heisenberg [32] and Schrödinger [67] published seminal papers on quantum theory in 1925 and 1926, and in 1929 Paul Dirac claimed that "the fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved" [16].

Efforts made to apply the formalism of quantum theory to chemical systems started as early as 1927, when Heitler and London [see manuscript letter in http://tinyurl.com/34dnkgr] developed the valence bond theory, a theoretical framework that is still in use. However, practical implementation of the principles of quantum chemistry had to wait for the development of computers. The first quantum mechanical calculations on model molecular systems were carried out in the 1950s, but the explosion of quantum chemistry took place during the 1970s, due to the combination of more powerful computers, efficient computational programs, and robust algorithms (for a review, see [66]). Since the 1970s, developments in quantum chemistry have paralleled improvements in computer architecture and the refinement of programs and algorithms. The Nobel Prize in Chemistry that was awarded in 1998 to John Pople (one of the leading scientists in translating quantum theories into efficient algorithms) can be viewed as the scientific community's recognition that quantum chemistry had come of age. Today, quantum chemistry affords a well defined framework for the rigorous treatment of most systems of interest for either organic or inorganic chemists, and very significantly contrib-

* Correspondence: M. Orozco, Departament de Bioquímica i Biología Molecular, Facultat de Química, Universitat de Barcelona, Martí i Franqués 1, E-08028 Barcelona, Catalonia, EU. Tel. +34-934037156. Fax +34-934037157. Email: modesto@mmb.pcb.ub.es

utes to the representation of systems on the frontier of biology and physics.

The impact of quantum chemistry on our understanding of the behavior of molecular systems can hardly be exaggerated. However, it would be unfair to ignore the contributions to theoretical chemistry arising from other methodological frameworks. It is worth noting, for instance, the efforts made by researchers such as E.J. Corey to rationalize organic synthesis from semi-empirical rules [9], and the coordinated efforts of different groups (e.g., Lifson, Allinger, and Hagler) to rationalize the conformational preferences of large molecules using classical mechanics [81].

In summary, theory has been part of the essence of chemistry since the early work of Dalton in the 18th century. In contrast, the origins of biology were quite different, being mostly linked to the work of scientists whose intensive efforts allowed the description and classification of nature, such as those of Linnaeus, in categorizing animal and vegetal kingdoms, and of van Leeuwenhoek, in the characterization of microbial organisms.

The intelligent observation of nature allowed the generation of phenomenological rules that, based on previous experience, could be applied to predict the evolution of biological systems. Examples of this type of reasoning are provided by the work of Darwin, Lamarck, and Mendel. Even now, much of the approach to the life sciences relies on this scientific paradigm.

The development of biochemistry, a hinge science between chemistry and biology that resulted from the work of researchers such as Pasteur, helped to narrow the conceptual gap between biology and chemistry. Biochemistry, for example, teaches us that complex biological processes such as metabolism are merely a well-coordinated and regulated network of simple chemical reactions favored by the presence of extremely efficient catalysts. However, the most important link in the bridge between biology and chemistry was the emergence, in the middle of the 20th century, of structural biology. Structural techniques made evident the chemical nature of biological models, thus facilitating the rationalization of biological processes using the same rules that have helped us to understand chemical systems.

The discovery of the structure of the DNA double helix by Watson and Crick in 1953 is still considered as the most relevant discovery in modern biology, since it demonstrated the possibility to gain insight into complex biological processes through basic chemical principles. In 1959, two other structural biologists, Kendrew and Perutz, solved the three-dimensional structure of the first protein [37,56], thus providing chemical clues regarding the functioning of these biological macromolecules. With these discoveries, the door for the atomistic interpretation of biology was opened, and theoretical chemistry, which had been evolving in parallel with structural biology, took advantage of the new approach.

The mainstream of research in biology is now dominated by automated high-throughput experiments that have been used to extract genomic information for most of the species of human interest. Indeed, with the latest generation of sequencing machines a human genome can be sequenced every week. The recently developed field of transcriptomics has enabled scientists to draw up an RNA expression map of tissues under normal vs. pathological conditions, and high-throughput proteomic techniques can identify all the proteins expressed at a given time in a particular tissue. Structural methods have also become of a high-throughput nature, allowing the structures of complex molecular machines, such as the ribosome, to be solved. At the time this review was being written, the Protein Data Bank (PDB) contained detailed structural information on nearly 60,000 proteins and more than 2000 nucleic acids, with the number of deposited structures growing exponentially. Overall, biology is now facing the challenge of managing the huge amount of data so that it can be readily accessed in attempts to understand the molecular basis of biological phenomena.

This review discusses how theoretical chemistry can help biology to rationalize all these data. For the sake of brevity, it focuses on specific areas in which theoretical chemistry has made relevant contributions to our knowledge of the principles that operate in biological systems.

Small molecules

In addition to macromolecules, small compounds are equally essential for cell life, as building blocks of macromolecular systems, whose properties are determined by the inherent features of the individual monomers, and because of their specific functions, which include signaling, allosterism, and mediators of metabolic pathways. Therefore, the study of simple nucleobases, amino acids, sugars, and other metabolites can provide very valuable information on the general properties of macromolecules and the molecular basis of biochemical processes.

The small size of these compounds permits the application of high levels of quantum mechanical (QM) theory to characterize their structural and chemical properties. A well-known example is the study of peptide bond isomerism. Most protein residues display a *trans* conformation around amide bonds with *cis*↔*trans* conversion hindered by a large energy barrier (around 20 kcal/mol). Under biological conditions, however, *cis*↔*trans* isomerism is largely favored for proline residues due to the catalytic action of rotamases. The molecular mechanism of *cis*↔*trans* isomerism is still unclear, as are the differential trends of the process in aqueous solution, apolar solvent, and the protein's interior. Finally, the molecular determinants of the kinetic efficiency achieved by rotamases remain to be fully elucidated. The combination of high-level quantum mechanical studies [39,40] with classical simulations [48] has helped us to define a mechanism such as that illustrated in Fig. 1. The free-energy barrier for isomerization in solution is around 19 kcal/mol and involves crossing both *syn* and *anti* transition states (Fig. 1). Desolvation reduces the barrier to 15 kcal/mol, favoring the *anti* transition state, which accordingly is around 3000-fold more populated than the *syn* transition state. This specific desolvation effect, combined with specific ion–dipole interactions, accounts for the enzymatic acceleration of the reaction.

Another case in which the study of small molecules provides value information on the behavior of large macromolecules is

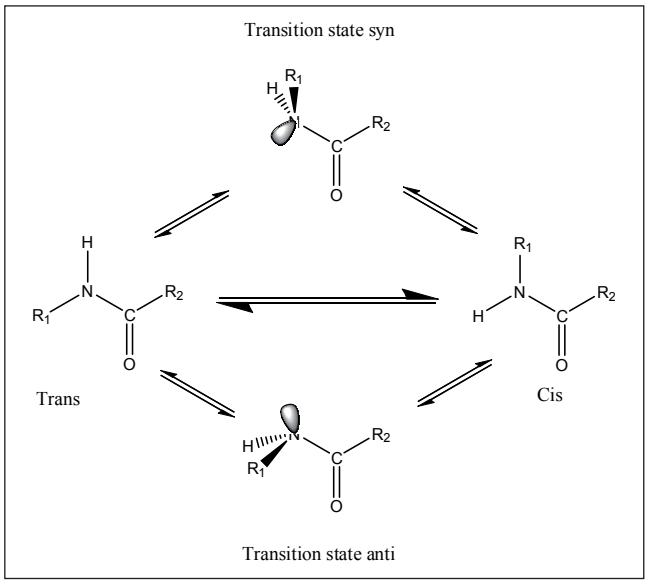

Fig. 1. Isomerization mechanism for amide bonds.

DNA, whose integrity relies on the formation of purine-pyrimidine hydrogen bonds, which in turn depends on the tautomeric state of the nucleobases. Already in their 1953 paper, Watson and Crick recognized the dependency of their model of A·T and G·C pairings on the prevalence of keto/amino tautomers with respect to the enol/imino forms [81], something that was in fact suggested to Watson and Crick by the theoretical chemist Max Delbrück, in contradiction of the experimental data available at that time. It is now clear that Delbrück's suggestion was correct and that G·C and A·T pairings are prevalent in DNA; in fact, they are responsible for the maintenance of the genetic code. Note that if the standard keto-amino tautomeric preference were altered, the recognition rules between nucleobases would also change, as shown in Fig. 2. This was suggested by Fresco as the major mechanism for spontaneous mutation [77].

This line of reasoning poses many exciting questions of major biological importance that can be solved with the aid of theoretical chemistry. For example, why are the keto-aminos the major species for coding nucleobases? Why do experimental observations suggest that tautomeric species present in the gas phase differ from those in solution? Is tautomerism modulated by the DNA environment or by modifications in the nucleobase? Can we understand mutagenic properties of modified bases by considering that they are in their minor tautomeric states? Or, even more interesting, can we design modified nucleotides with tautomerism-driven dual recognition patterns?

A combination of QM calculations, self-consistent reaction field calculations, molecular dynamics (MD), and statistical mechanics calculations (see below) has guided our understanding of the tautomeric scenario of different nucleobases in a variety of environments. The most interesting case is cytosine, which changes its tautomeric state between apolar environments such as the gas phase and water [8], and which also can change in specific DNA environments [60,72]. The same type of calculations have led to the detection of unusual nucleobases with dual tautomeric states and displaying different recognition patterns [2] according to the DNA environment, thus paving the way to important applications in biotechnology. Furthermore, analyses of tautomeric scenarios of modified nucleosides have been crucial to our understanding of the mutagenic properties of different DNA lesions (see examples in [19,34,35,49,79]).

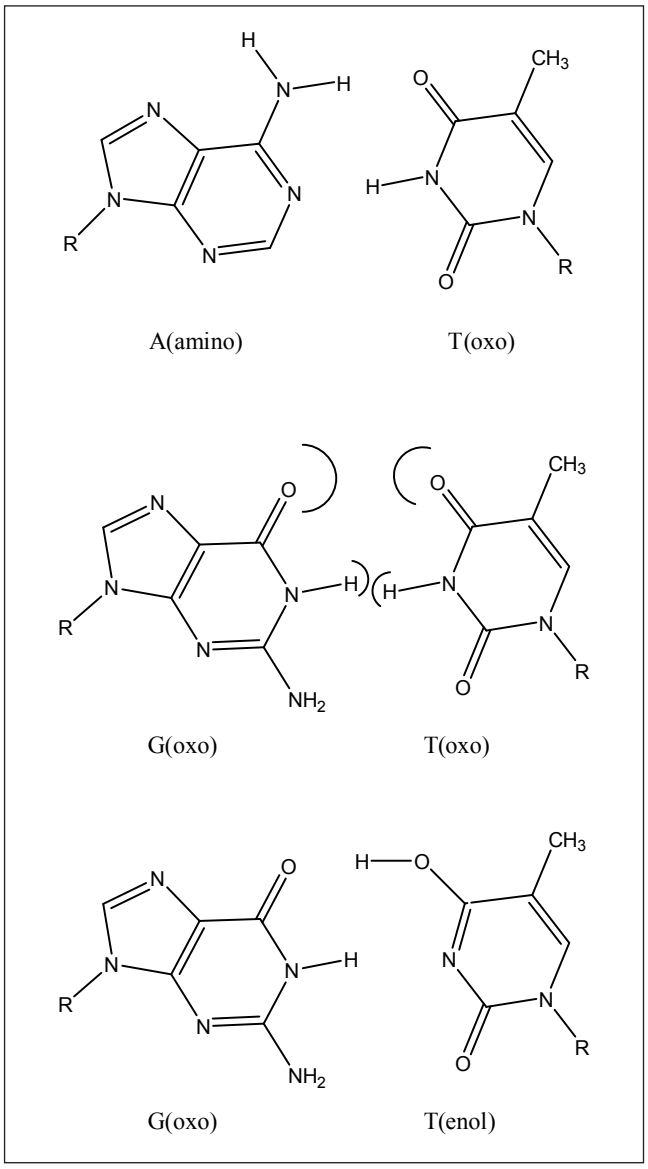

Fig. 2. Example of how tautomeric changes in one nucleobase (thymine in this case) can force the A→G transversion in DNA (bottom pair), which is prevented for the canonical tautomer (top and middle pairs).

QSAR and chemogenomics

The work of Hansch and others during the 1970s and 1980s demonstrated that it was possible to relate the molecular descriptors of a series of molecules with their biological activities following the "extrathermodynamic approach" (Eq. (1), [22,29]).

$$\log A = \Im(X_1, \ldots Xn) \quad (1)$$

where A is the biological activity, *X1…Xn* are molecular descriptors, and $\Im$ is an a priori unknown function.

The use of Eq. (1) does not require prior knowledge of the structure of the biological receptor. Rather, the biological activities determined for a series of compounds are used to establish a numerical formalism for $\Im$, which can then be applied to predict the activity of other derivatives.

Equation (1) can be used to identify quantitative structure-activity relationships (QSAR), which have an enormous impact in guiding drug design [22,29]. Although the first QSAR studies were based on experimentally determined molecular descriptors (such as partition coefficient, molecular refractivity, and molecular dipole), most QSAR descriptors are derived from theoretical chemistry [59], and there are international assessments that allow the accuracy of these theoretically derived descriptors to be evaluated [74].

The advantages of theoretical methods with respect to experiments aimed at deriving QSAR descriptors are two-fold: (i) they are easy to obtain even before the synthesis of the compound, and (ii) they easily capture three-dimensional information, which is often difficult to obtain from experiments. An example of the latter point is seen in Fig. 3, which illustrates the richness of the information on the hydrophobic/hydrophilic properties of a putative drug compared with the flat information provided by inspection of its partition coefficient.

One of the possibilities that theory offers to researchers is quantification of the similarity/diversity of molecules [5,36]. This information, which can be obtained using different levels of theory, has been extremely beneficial in the design of candidate drugs based on their similarities with known active drugs. Similarity techniques now form the core of chemogenomics, in which theoretical chemistry is combined with the tools of bioinformatics to place drugs in the context of gene relationships in order to detect cross-interactions (found by looking for drugs known to act on target A that are similar to drugs acting on target B), or to design "dirty" drugs able to tackle simultaneously different targets in a network, thus increasing the likelihood of a favorable pharmacological response [30].

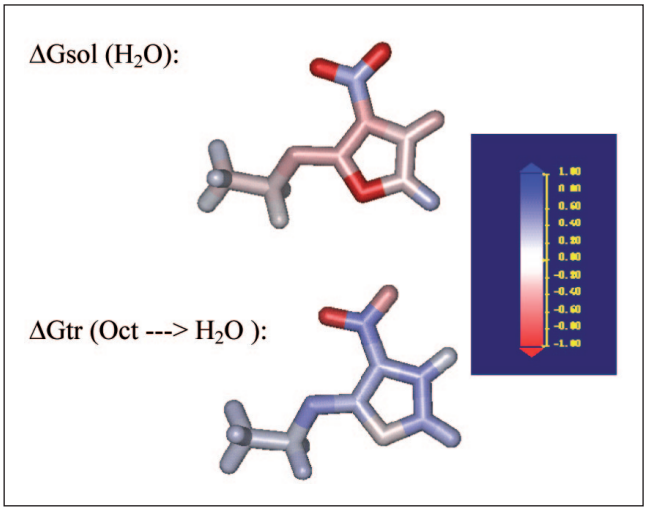

Fig. 3. Fractional contributions to hydration free energy and octanol→water partition coefficient obtained by using a theoretical self-consistent reaction field method [41,73]. The plot illustrates all the details of the spatial distribution of hydrophobic and hydrophilic regions in the molecule.

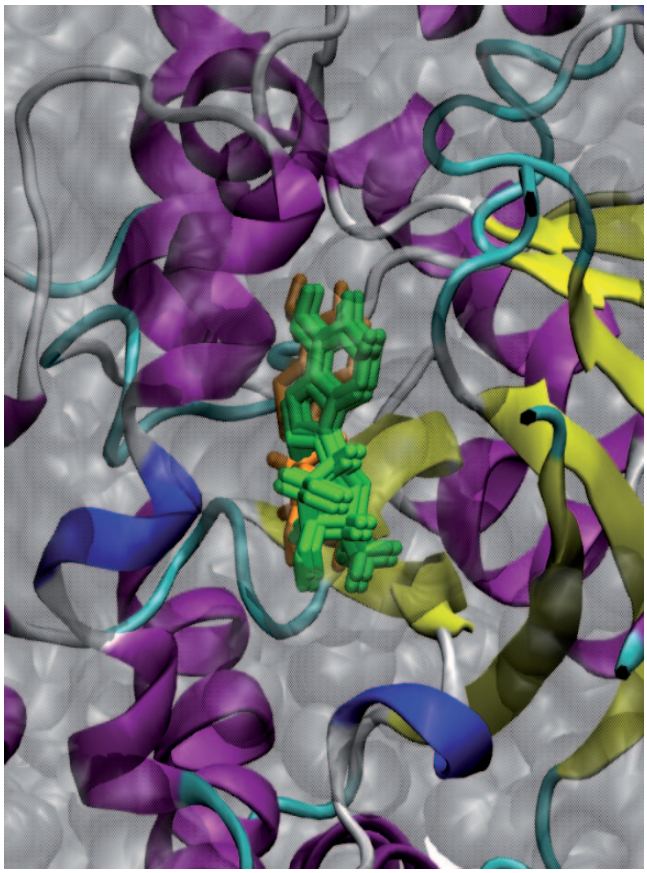

Fig. 4. Example of docking results obtained using our CMIP program [23] on adenosine deaminase. The optimal poses of the docked ligand (1-deaza-adenosine) are shown in green, while the X-ray determined binding mode is shown in orange.

Structure-based drug design

As described above, ligand-based approaches afford a conceptual framework to gain insight into the relationships between a drug's structure and its activity. However, knowledge of the three-dimensional structure of the drug's receptor greatly increases the probability of success, since docking techniques [10,43] can be used to screen chemical databases for molecules that fit well in the binding pocket of proteins.

Current docking methods use fast sampling techniques and simple potential functions to scan the potential binding modes of drugs to the target protein. The proposed binding modes ("poses" in docking jargon) are then scored using more complex semi-empirical functions that have been refined by training with known ligand databases [43]. The final output of docking algorithms is a set of drugs and potential binding modes (Fig. 4) that are labeled by estimating the theoretical free energy of binding. Docking methods are designed to enable efficient screenings of libraries containing millions of drugs. Despite problems in representing induced-fit structural changes in the protein [10], they highly improve (by a factor of 100) the chances of finding effective ligands. However, while docking algorithms are the best alternative to experimental high-throughput methods in searching for new leads in drug design, they are not accurate enough for lead-optimization, in which higher-level calculations (see below) are required.

Protein function

An understanding of how proteins function is one of the main objectives of theoretical chemists working in biological systems. Some proteins perform mechanical work much more efficiently than any human-made engine. Many can recognize ligands with nanomolar to even femtomolar binding affinities that are capable of triggering broad-ranging biological responses. Other proteins bind and carry ligands across long distances, delivering them to specific sites. Finally, there are proteins that act as catalysts of complex reactions, achieving much larger enhancements of the reaction rate than possible with any human-made catalyst. The theoretical study of protein function presents two major challenges. First, there is the need to represent the dynamic behavior of proteins, and second, to account, in some cases, for chemical reactions.

The study of protein dynamics is complex, but can be achieved by means of two main approaches [18], coarse-grained methods and atomistic methods. Within the coarse-grained approach, the protein is represented by a series of beads (typically localized at Cα) that interact with each other by means of simple harmonic potentials, such as those defined in Eq. (2).

$$E = \sum_{i,j} \Gamma_{ij} \left(d_{ij} - d_{ij}^{0}\right)^2 \tag{2}$$

where Γ_{ij} is a distance-dependent or delta function equal to 1 if beads i and j are close in space and to 0 otherwise, d_{ij} is the actual distance between beads, and d_{ij}^{0} the optimum distance as determined from the experimental structure.

The movements of a protein subjected to potential energy in Eq. (2) can be obtained by integration of the corresponding Langevin equations of motion [18], as shown in Eq. (3).

$$m_i \vec{a}_i = -\gamma \vec{\nu}_i + \vec{F} + \eta(t) \tag{3}$$

where m is the mass, $\vec{a}$ the acceleration, $\vec{\nu}$ the velocity, and $\vec{F}$ the force (computed as $\vec{F}_i = -dE/dr_i$) acting on atom i; γ is a friction coefficient and the stochastic term $\eta_i(t)$ is computed using Gaussian white-noise functions (see [35] for details).

Alternatively, within the normal mode analysis (NMA) framework, protein movement can be deduced from the set of eigenvectors and eigenvalues obtained by diagonalization of the Hessian matrix:

$$H_{ij} = \left(\frac{\partial E^2}{\partial x_i x_j}\right) \tag{4}$$

A last option for the coarse-grained representation of protein movements consists of mapping harmonic potentials as square potentials. This allows the derivation of trajectories by using ballistic equations of motion, thus avoiding the computationally expensive calculations implicit in the integration of Langevin equations of motions or in the diagonalization of the Hessian's matrix.

Coarse-grained calculations are very efficient from a computational point of view and provide results of surprisingly great accuracy [4,18]. Unfortunately, they do not yield the atomistic details that are often important for understanding protein function. This shortcoming can be corrected by atomistic MD simulations, in which the protein is represented at atomic resolution as being surrounded by thousands of water molecules and ions, i.e., mimicking physiological conditions. MD simulations are coupled to highly detailed potential functions (force-fields) that have been parametrized to reproduce high-level QM and experimental data [3]. Trajectories for individual atoms are collected by numerical integration of Newton's equations of motion over very small time intervals (*dt*; typically femtoseconds):

$$m_i \vec{a}_i = -dE/dr_i \tag{5}$$

$$\vec{\nu}_i(t) = \vec{\nu}_i(t=0) + \int_{t=0}^{t=dt} \vec{a}_i(t)dt \tag{6}$$

$$\vec{r}_i(t) = \vec{r}_i(t=0) + \int_{t=0}^{t=dt} \vec{a}_i(t)dt \tag{7}$$

MD simulations provide a Boltzmann ensemble of the protein and its solvent environment (Fig. 5) that can be used to reproduce any experimental observable. Additionally, MD can reproduce spontaneous or externally induced conformational changes, thereby allowing studies of the mechanical work generated by proteins, the nature of allosteric effects, and the mechanisms of ligand-induced fit [38,39].

Despite the impressive power of MD simulations, the use of classical force-fields and fixed molecular topologies impedes their application in the study of chemical reactions occurring at the active sites of enzymes. In these cases, theoretical studies require a hybrid potential-energy function (the hybrid Hamiltonian) representing, at the QM level, interactions at the active site (*as*), whereas external interactions rely on a more efficient classical formalism (MM; see [68,69] for additional details and examples of use).

$$\hat{H} = \left\{\hat{H}_{QM}\right\}_{i \in as} + \left\{E_{MM}\right\}_{i \notin as} + \left\{E_{QM/MM}\right\}_{i \notin as, j \in as} \tag{8}$$

The hybrid Hamiltonian outlined in Eq. (8) can be implemented with few changes in the MD formalism, but the cost of the QM part of the calculation makes the calculation highly expen-

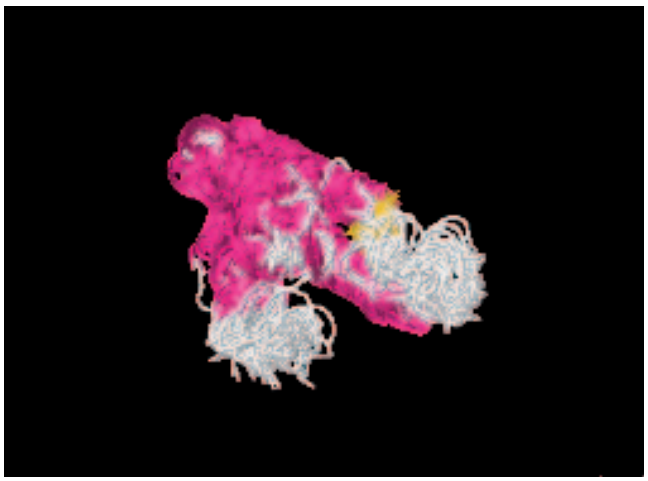

Fig. 5. Ensemble of conformations representing the equilibrium dynamics of apocytocrhome B562 in aqueous solution. Image taken from our webserver [http://mmb.pcb.ub.es/MoDEL].

sive. Thus, while current classical MD of proteins are reaching the 100 ns to 1 μs time scale, QM-MM calculations are typically performed in the multi-picosecond to nanosecond time scale [68,69].

Prediction of protein structure

X-ray and nuclear magnetic resonance (NMR) are dramatically increasing the number of proteins of known structure. As noted above, the 2010 version of the PDB contains around 60,000 entries. Nevertheless, only around 40% correspond to unique proteins and only 15% of these are human proteins. In contrast, around 93,000 human proteins are expected based on sequence data, probably nearly 200,000 if splicing variants are considered [7,58]. Thus, the PDB currently covers less than 10% of human proteins (probably even less than 2%), underlining the need to obtain theoretical models of protein structure in order to understand the function of proteins and to design more effective drugs.

Interest in predicting protein structure has dominated theoretical work in biochemistry for more than half a century, since the sequence of a protein was demonstrated to contain all the information needed to achieve its three-dimensional structure. In principle, the prediction of a protein's structure should imply a simple free-energy optimization process, which could be solved using long MD simulations coupled to a classical force-field. The problems of this pure force approach are twofold: (i) there is no guarantee that physical force-fields, parametrized to correctly reproduce native folds, will also properly reproduce the unfolded structures, and (ii) folding is a statistical process resulting from an ensemble of trajectories, typically on a millisecond to second time scale.

As mentioned above, recent improvements in computers and software are now making it feasible to run microsecond-long MD simulations. Furthermore, in the coming years, next-generation supercomputers, hybrid CPU-GPU systems, and especially special-purpose computers are expected to facilitate millisecond-long trajectories, thus enhancing the possibility to perform real folding simulations based on physical potentials derived from "first principles" calculations. This, in turn, has well evidenced the need for force-field refinement, currently the focus of intense efforts by many research groups [47].

In the absence of a final answer to the folding problem derived from rigorous theoretical calculations, more empirical approaches are being used, and with notable success. Methods for the prediction of secondary structure have now reached maturity, guaranteeing success rates above 70% for soluble proteins. More challenging is the prediction of three-dimensional structure, since this is defined by a myriad of weak interactions acting cooperatively to stabilize a protein's native form.

Current methods for protein prediction are based on two fundamental paradigms: (i) structure is more conserved than sequence, i.e., proteins with similar sequence display very similar structures, and (ii) the folded structure of a protein is its most stable conformation. The standard procedure for structure prediction first determines whether the database contains homologous proteins of known structure. If there are homologues with a sequence identity above 30%, comparative modeling techniques provide excellent structures for the problem protein by using the known structure as template [65].

The impact of comparative modeling in biology is enormous and is expected to increase as the number of template structures in the PDB increases. Recent calculations by our group demonstrated that good homology models can be built for nearly 55% of the known human proteins [45]. However, there are still many proteins (human or belonging to other species) for which there is no clear homologue and, accordingly, they are not targets for comparative modeling. In many of these cases, fold recognition techniques can be used and structural models can be derived by employing semi-empirical scoring functions that evaluate the ability of a protein to adopt a known fold [27]. In the remaining cases, when neither comparative modeling nor fold recognition techniques are applicable, the *ab initio* folding approach can be used. Within this paradigm, the protein is folded by means of a Monte Carlo strategy or using other sampling techniques that optimize contacts, as defined by knowledge-based potentials and fragment libraries [21].

The protein prediction community is large and very well organized. This has facilitated the organization of periodic assessments (critical assessment of structural prediction, CASP), in which the abilities of current methods to determine the structure of proteins experimentally known but not yet deposited in the PDB are analyzed [80]. CASP experiments have fueled the refinement of predictive methods, and publication of the results has provided the general user community with a very good guide regarding the expected quality of the structural models. The spirit implicit in CASP experiments highly favors other areas of theoretical chemistry, in which the expected accuracy of the calculations is not always evident.

Protein interactions and networks

Recent large-scale experiments [1,64] have illustrated that many cellular processes are guided by the formation of transient protein-protein complexes, whose structures are typically difficult to determine by high-resolution experimental methods but instead require the use of simulation tools. A variety of methods based on protein evolutionary information [78], the optimization of empirical functions or physical potentials [6,20], and a combination of the two [17,28] have been proposed. Current state-of-the-art methods can reasonably trace regions of proteins that are probably involved in protein-protein contacts. Structural models of the protein complexes are of good quality when dimerization does not introduce major changes in monomer geometry, but they are potentially rather inaccurate when the interaction triggers large structural deformations [57]. Protein docking methods are periodically analyzed in a CASP-like blind experiment named CAPRI (critical assessment of prediction of interactions), which has illustrated the continuous increase in the performance of these methods [see website information at http://mmb.pcb.ub.es/capri2009].

Recent trends in the area of protein-protein docking are: (i) to introduce coarse-grained estimates of protein flexibility and (ii) to enrich theoretical predictions by using low-resolution experimental data, such as site-directed mutagenesis [17], mass spectrometry, hydrodynamic measures of molecular shape, or small-angle scattering spectra [57], which help to re-score theoretical models, rejecting potential structures that seem acceptable for current scoring functions [15].

Protein-protein docking methods are quite efficient and they allow proteome-scale simulations [44]. This provides the opportunity to complete those interactomes outlined in large-scale genomic and proteomic experiments and to generate structural models for already-known complexes, complementing experimental information that is currently derived from techniques such as electron tomography (see [1,64] for a discussion). Thus, protein-protein docking simulations serve as a bridge between structural bioinformatics techniques directly derived from computational chemistry and systems biology. As a result, it should soon be possible to use theoretical chemistry to design a small drug able to bind a given protein target and, with related techniques, to predict the impact of this interaction on the entire cellular metabolome or interactome.

Nucleic acid simulations

Nucleic acids encode the primary sequence of proteins and contribute to the regulation of gene expression, including during the synthesis of proteins from RNA on ribosomes. They are structurally less complex than proteins since they are formed by only four different building blocks. The secondary structure of nucleic acids consists of regular double-stranded helices, defined in most cases by A·T and G·C pairs. However, despite this apparent simplicity, the theoretical study of nucleic acids is more difficult than that of proteins, for the following reasons: (i) the main intrinsic stabilizing term in the structure of nucleic acids is dispersion, whose theoretical representation is especially difficult, (ii) they are extremely flexible, and (iii) solvent and ion effects (always complex to represent since they are long-range) play crucial roles in determining the fine details of the three-dimensional structure.

Quantum simulations of nucleic acids require the inclusion of electron correlation, which dramatically increases computational costs. Accordingly, QM calculations have been limited to the nucleotide (or di-nucleotide) level. High-level QM studies have helped us to understand the physics of nucleobase-nucleobase interactions in the gas phase or under pure solvent conditions [76] and to discuss the relative stability of different tautomers (see above). They also have been crucial in the parametrization of the force-field for classical simulations [53]. However, the study of complex nucleic acid structures is far beyond the capabilities of quantum chemistry (and will remain so in the next few years) such that most simulations on these systems have been carried out at the classical level.

The leading technique in the study of medium-sized nucleic acids (from 4 to 140 base pairs) is MD simulations. Recent improvements in force-fields [53] and in simulation conditions [50] combined with the use of new-generation supercomputers have made it possible to study the structure and dynamics of small helices in the micro-second range [52], which is the time frame for certain key biophysical processes such as chemical unfolding or local folding [55]. Indeed, the tremendous power of MD simulations of nucleic acid reactions has allowed the study of the effect of mutations, in the form of chemically modified bases, on DNA structure [11,12,24,75], the physical properties of unusual nucleic acids [70,71], nucleic acids under atypical conditions [61,63], the nature of nucleic acid transitions [46], the basis for DNA–drug recognition [31], the effect of ions on nucleic acid structure [62], folding/unfolding [13], and even the characterization of new structures of nucleic acids prior to their experimental determination [13,14].

However, classical atomistic MD is still limited to the study of systems comprising only a few hundred nucleobases [50,51], far from the chromatin-scale. This has encouraged the development of macroscopic models that treat the DNA structure as elastic rods [50] and do not explicitly include sequence effects. These macroscopic models can be used to gain general information on the average properties of DNA, as in the study of large plasmids or the analysis of DNA packing in viruses [50]. When sequence effects are important but the systems are too large for atomistic descriptions, mesoscocopic methods are needed. These methods represent nucleic acids at the base-pair level and assume that conformational changes can be expressed as combinations of three rotations (tilt, roll, twist) and three translations (rise, slide, shift) of one base pair with respect to the other. In terms of the harmonic approach, this means that the energy of a given regular duplex is defined as:

$$E = \Xi(\Delta X)^2 \tag{9}$$

where ΔX is the deformation, and Ξ the stiffness matrix:

$$\Xi = \begin{pmatrix} K_w & K_{wr} & K_{wt} & K_{ws} & K_{wl} & K_{wf} \\ K_{wr} & K_r & K_{rt} & K_{rs} & K_{rl} & K_{rf} \\ K_{wt} & K_{rt} & K_t & K_{st} & K_{tl} & K_{tf} \\ K_{ws} & K_{rs} & K_{st} & K_s & K_{ls} & K_{lf} \\ K_{wl} & K_{rl} & K_{tl} & K_{ls} & K_l & K_{lf} \\ K_{wv} & K_{rf} & K_{tf} & K_{lf} & K_{lf} & K_f \end{pmatrix} \tag{10}$$

where diagonal terms account for stiffness associated with pure twist (w), roll (r),tilt (t), slide (s), slide (l), and shift (f) deformations, and non-diagonal terms represent cross-interactions.

Practical determination of the stiffness constants is done by taking advantage of the Einstein equation applied to MD ensembles of equilibrium trajectories [50], as shown in Eq. (11).

$$\Xi = k_B T \subset^{-1} \tag{11}$$

where $\subset$ is the covariance matrix obtained by projecting the atomistic MD into the nucleobase-pair helical space.

Efforts have been made by the theoretical chemistry community to characterize the equilibrium and elastic properties of short fragments of DNA. This information will be valuable in re-creating the structure and elastic properties of long DNA fragments [38,54] using modified nearest-neighbor models, which

can characterize the physical properties of DNA at the genomic scale [25,26]. For example, our group has found correlations between DNA regulatory regions and unusual physical properties, thus offering a method able to predict the positioning of transcription factors with excellent accuracy [25].

Final remarks

Coulson's remark that "living organisms are the most perverse of all chemical systems" illustrates well the complexity of biological systems when studied from the viewpoint of theoretical chemistry. Biological systems are very large, diffuse, and have been inadequately described, with data that are in many cases poor in information. Even worse, the nature of the biological problem is often not well defined, hampering its formulation with the clarity needed for theoretical calculations. Despite these shortcomings, the recent rise of theoretical chemistry as a major rationalizing tool for biology has been impressive. New theoretical developments, accessibility to more powerful computers, and the availability of more accurate biological data are, together, opening new doors, in which theoretical and computational chemistry can guide biologists in their characterization of the chemical bases of living organisms, from the small molecular detail of interactions to the global description of entire ecosystems.

References

[1] Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. Nature Biotechnol 22:1317-1321

[2] Blas JR, Luque FJ, Orozco M (2004) Unique tautomeric properties of isoguanine. J Am Chem Soc 126:154-164

[3] Brooks III CL, Karplus M, et al. (1987) Proteins: A theoretical Perspective of Dynamics, Structure and Thermodynamics. Cambridge University Press, Cambridge

[4] Camps J, Carrillo O, Emperador A, et al. (2009) FlexServ: an integrated tool for the analysis of protein flexibility. Bioinformatics 25(13):1709-1710

[5] Carbó-Dorca R, Mezey PG (2001) Fundamentals of Molecular Similarity. Springer, New York

[6] Cheng T, Blundell TL, Fernández-Recio J (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 68:503-515

[7] Clark F, Thanaraj TA (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum Mol Genet 11:451-464

[8] Colominas C, Luque FJ, Orozco M (1996) Tautomerism and protonation of guanine and cytosine. Implications in the formation of triplex DNA. J Am Chem Soc 118: 6811-6821

[9] Corey EJ, Wipke T (1969) Computer-assisted design of complex organic synthesis. Science 166:178-192

[10] Cozzini P, Kellogg GE, Spyrakis F, et al. (2008) Target flexibility: an emerging consideration in drug discovery and design. J Med Chem 51:6237-6255

[11] Cubero E, Laughton CA, Luque FJ, Orozco M (2000) Molecular dynamics study of oligonucleotides containing difluorotoluene. J Am Chem Soc 122:6891-6899

[12] Cubero E, Güimil-García R, Luque FJ, Eritja R, Orozco M (2001) The effect of amino groups on the stability of DNA duplexes and triplexes based on purines derived from inosine. Nucleic Acid Res 29:2522-2534

[13] Cubero E, Luque FJ, Orozco M (2001) Theoretical studies of d(A:T)-based parallel-stranded DNA duplexes. J Am Chem Soc 123:12018-12025

[14] Cubero E, Abrescia N, Subirana JA, Luque FJ, Orozco M (2003) Theoretical study of a new structure of DNA: The antiparallel Hoogsteen duplex. J Am Chem Soc 125:14603-14612

[15] D'Abramo M, Meyer T, Bernadó P, Fernandez-Recio J, Orozco M (2009) On the use of low resolution data to improve structure predictions of proteins and protein complexes. J Chem Theor Comput 5:3129-3137

[16] Dirac PAM (1929) Quantum mechanics of many-electron systems. Proc R Soc London Ser A 123:714-733

[17] Dominguez C, Boelens R, Bonvin AM (2003) Haddock: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731-1737

[18] Emperador A, Carrillo O, Rueda M, Orozco M (2008) Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. Biophys J 95:2127-2138

[19] Faustino I, Aviño A, Marchán I, Luque FJ, Eritja R, Orozco M (2009) The Unique tautomeric and recognition properties of thio-thymines? J Am Chem Soc 131:12845-12853

[20] Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. Proteins 58:134-143

[21] Fetrow JS, Skolnick J (1998) Methods for prediction of protein function from sequence using the sequence-to-structure-to-function parading with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 281:949-9[22] Franke R (1984) Theoretical Drug Design Methods. Elsevier, Amsterdam

[23] Gelpí JL, Kalko SG, Barril X, Cirera J, de La Cruz X, Luque FJ, Orozco M (2001) Classical molecular interaction potentials: improved setup procedure molecular dynamics simulations of proteins. Proteins 45(4):428-437

[24] Gómez-Pinto I, Cubero E, Kalko SG, Monaco V, Van der Marel G, van Boom JH, Orozco M, González C (2004) Effect of bulky lesions on DNA: Solution structure of a DNA duplex containing a cholesterol adduct. J Biol Chem 279:24552-24560

[25] Goñi JR, Pérez A, Torrents D, Orozco M (2007) Studying the role of DNA physical properties on gene regulatory mechanisms in vertebrates. Gen Biol 8:R263

[26] Goñi JR, Fenollosa C, Pérez A, Torrents D, Orozco M (2008) DNAlive: A tool for the physical analysis of DNA at the genomic scale. Bioinformatics 95:1731-173

[27] Gozki A (2003) Fold Recognition Methods. In: Bourne PE,

Weissig H (eds) Structural Bioinformatics. Wiley-Liss, New Jersey

[28] Gray JJ, Moughan SE, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D (2003) Protein-Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. J Mol Biol 331(1):281-299

[29] Hansch C (1973) Quantitative structure-activity relationships in drug design. In: Cavallito CJ (ed) Structure-Activity Relationships. Pergamon Press, Oxford

[30] Harris CJ, Stevens AP (2006) Chemogenomics: structuting the drug discovery process to gene families. Drug Discov Today 11:880-888

[31] Harris SA, Gavathiotis E, Searle MS, Orozco M, Laughton CA (2001) Co-operativity in drug-DNA recognition: a molecular dynamics study. J Am Chem Soc 123:12658-12663

[32] Heisenberg W (1925) Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen. Z Phys 33:879-893

[33] Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature 450:913-916

[34] Hernández B, Luque FJ, Orozco M (1996) Tautomerism of Xanthine Oxidase Subtrates Hypoxathine and Allopurinol. J Org Chem 61:5964-5971

[35] Hernández B, Soliva R, Luque FJ, Orozco M (2000) Misincorporation of 2'-deoxyoxanosine into DNA: A molecular basis for NO-induced mutagenesis derived from theoretical calculations. Nucl Acid Res 28:4873-4883

[36] Johnson MA, Maggiora GM (eds) (1990) Concepts and Applications of Molecular Similarity. Wiley, Amsterdam

[37] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181:662-666

[38] Lavery R, Zakrzewska K, Beveridge D, et al. (2010) A systematic molecular dynamics study of the nearest-neighbor effects on base pair and base step conformations and fluctuations in B-DNA. Nucleic Acids Res 38:299-313

[39] Luque FJ, Orozco M (1993) Ab Initio study of the reactivity of planar and twisted amides. New insights into the amide resonance. J Chem Soc Perkin Trans II 2:683-690

[40] Luque FJ, Orozco M (1993) Theoretical study of N-methylacetaminde in vacuum and aqueous solution: Implications for the peptide bond isomerisation. J Org Chem 58:63975-6405

[41] Luque FJ, Barril X, Orozco M (1999) A new method for the fractional description of solvation free energies. Application in drug-design. J Comp-Aided Mol Design 13:139-153

[42] Ma J, Karplus M (1998) The allosteric mechanism of the chaperonin GroEL: a dynamic analysis. Proc Natl Acad Sci USA 95:8502-8507

[43] Mohan V, Gibbs AC, Cummings MD, Jaeger EP, Desjarlais RL (2005) Docking: successes and challenges. Curr Pharm Des 11:323-333

[44] Mosca R, Pons C, Fernández-Recio J, Aloy P (2009) Pushing Structural Information into the Yeast Interactome by High-Throughput Protein Docking Experiments. PLoS Comput Biol 5(8): e1000490

[45] Novoa E, Ribas de Pouplana LI, Barril X, Orozco M. To be published

[46] Noy A, Pérez A, Laughton C, Orozco M (2007) Theoretical study of large conformational transitions in DNA: The B↔A conformational change in water and ethanol/water. Nucleic Acid Res 35:3330-3338

[47] Onufriev M, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation program. J Comput Chem 26:1668

[48] Orozco M, Tirado-Rives J, Jorgensen WL (1993) Mechanism for the rotamase activity of FK506 binding protein from MD simulations. Biochemistry US 32:12864-12874

[49] Orozco M, Hernández B, Luque FJ (1998) Tautomerism of 1-Me-uracil, 1-Me-thymine, and 1-Me-5-Br-uracil. Is tautomerism the reason for the mutagenicity of 5-Br-uridine? J Phys Chem B 102:5228-5233

[50] Orozco M, Pérez A, Noy A, Luque FJ (2003) Theoretical methods for the simulation of nucleic acids. Chem Soc Rev 32:350-364

[51] Orozco M, Noy A, Pérez A (2008) Recent advances in the study of nucleic acids flexibility by molecular dynamics. Curr Op Struct Biol 18:185-193

[52] Pérez A, Luque FJ, Orozco M (2007) Entering molecular dynamics in the biological time scale. Microsecond simulation of DNA. J Am Chem Soc 129:14739-14745

[53] Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE, Laughton CA, Orozco M (2007) Refinement of the AMBER force-field for nucleic acid simulations. Improving the representation of α/γ conformations. Biophysical J 92:3817-3829

[54] Pérez A, Lankas F, Luque FJ, Orozco M (2008) Towards a consensus view of B-DNA flexibility. Nucleic Acids Res 36:2379-2394

[55] Pérez A, Orozco M (2010) Real time atomistic description of DNA unfolding. Angew Chem Int Ed Eng 49:4805-4808

[56] Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT (1960) Structure of Haemoglobin. Nature 185:416-422

[57] Pons C, Grosdidier S, Solernou A, Pérez-Cano L, Fernández-Recio J (2010) Present and future challenges and limitations in protein-protein docking. Proteins 78:95-108

[58] Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:501-504

[59] Richards WG (1983) Quantum Pharmacology. Butterworhts, London

[60] Rueda M, Luque FJ, López JM, Orozco M (2001) Amino-imino tautomerism in derivatives of cytosine: effects on hydrogen bonding and stacking properties. J Phys Chem A 105:6575-6580

[61] Rueda M, Kalko S, Luque FJ, Orozco M (2003) The structure and dynamics of DNA in the gas phase. J Am Chem Soc 125:8007-8014
[62] Rueda M, Cubero E, Laughton CA, Orozco M (2004) Exploring the counterion atmosphere around DNA. What can be learnt from molecular dynamics simulations? Biophys J 87:800-811
[63] Rueda M, Luque FJ, Orozco M (2006) G-DNA can maintain its structure in the gas phase. J Am Chem Soc 128:3608-3619
[64] Russell RB, Aloy P (2008) Targeting and tinkering with interaction networks. Nature Chem Biol 4:666-673
[65] Sali A, Blundell TL (2003) Comparative Modelling by satisfaction of spatial restraints. J Mol Biol 234:779-815
[66] Schaefer HF (1984) Quantum Chemistry. Clarendon Press, Oxford
[67] Schrödinger E (1926) Quantisierung als Eigenwertproblem (Erste Mitteilung). Ann Phys 79:361-376
[68] Senn HM, Thiel W (2007) QM/MM Methods for Biological Systems. Top Curr Chem 268:173-290
[69] Senn HM, Thiel W (2009) QM/MM Methods for Biomolecular Systems. Angew Chem Int Ed 48:1198-1229
[70] Shields G, Laughton CA, Orozco M (1997) Molecular dynamics simulations of the d(T.A.T) triple helix. J Am Chem Soc 119:7463-7469
[71] Shields G, Laughton C, Orozco M (1998) Molecular Dynamic simulations of the PNA.DNA.PNA triple helix in aqueous solution. J Am Chem Soc 120:5895-5904
[72] Soliva R, Luque FJ, Orozco M (1999) Can G·C Hoogsteen-wobble pair contribute to the stability of d(G·C·C) triplexes? Nucleic Acid Research 27:2248-2255
[73] Soteras I, Morreale A, López-Bes JM, Orozco M, Luque FJ (2004) Group contributions to the solvation free energy from MST continuum calculations. Braz J Phys 34:48-57
[74] Soteras I, Forti F, Orozco M, Luque FJ (2009) Performance of the IEG-MST solvation continuum mode in a blind test prediction of hydration free energies. J Phys Chem B 113:9330-9334
[75] Spackova N, Cubero E, Sponer J, Orozco MJ (2004) Theoretical study of the guanine→6-thioguanine substitution in duplexes, triplexes and tetraplexes. J Am Chem Soc 126:14642-14650
[76] Sponer J, Jurecka P, Marchan I, Luque FJ, Orozco M, Hobza P (2006) Nature of base stacking. Reference quantum chemical stacking energies in ten unique B-DNA base pair steps. Chem Eur Journal 12:2854-2865
[77] Topal MD, Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. Nature 263:285-289
[78] Valencia A, Pazos F (2003) Prediction of protein-protein interactions from evolutionary information. In: Bourne PE, Weising H (eds) Structural Bioinformatics (chapter 20). Wiley-Liss, New Jersey
[79] Vázquez-Mayagoita A, Huertas O, Brancolini G, Sumpter G.B, Orozco M, Luque FJ, di Felici R, Fuentes-Cabrera M (2009) Ab initio study of structural, tautomeric, pairing and electronic properties of seleno-derivatives of thymine. J Phys Chem B 113:14465-14472
[80] Venclovas ZA, Fidelis K, Moult J (2001) Comparison of performance in succesive CASP experiments. Proteins 45(Suppl 5):163-170
[81] Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171:737-738

About the authors

Modesto Orozco received his Ph.D. in Chemistry from the Autonomous University of Barcelona in 1990. He is full professor of Biochemistry and Molecular Biology at the Department of Biochemistry at the University of Barcelona, principal investigator at the Institute of Biomedical Research (IRB), director of the structural bioinformatics section at the National Institute of Bioinformatics (INB), director of the Life Sciences Department at the Barcelona Supercomputing Center (BSC), and director of the joint IRB-BSC program on Computational Biology. He has worked in the Institute of Cancer Research (UK) and in the Chemistry Department of Yale University and is a recipient of several national and international awards. His main scientific interests are the theoretical study of chemical reactivity in condensed phases, the relationship between structure, dynamics, and function in biomolecules, and the development of methods for structure-based drug discovery.

F. Javier Luque received his Ph.D. in Chemistry from the Autonomous University of Barcelona in 1989. He is full professor in Physical Chemistry at the Department of Chemical Physics at the University of Barcelona. He is also head of the Computational Biology and Drug design group at the Institute of Biomedicine at the University of Barcelona. He has worked at the ETH (Switzerland), Pisa (Italy), and at Nancy (France), where he was Guest Professor. He is a recipient of the Autonomous Government of Catalonia's Distinguished Award to Young Researchers. His main scientific interests are the theoretical study of chemical reactivity in condensed phases, the relationship between structure, dynamics and function in biomolecules, and the development of methods for structure-based drug discovery.