Chimia 59 (2005) 550–553

ISSN 0009–4293

Computational Molecular Science for the Nutritional Industry#

Martin G. Grigorov*

Abstract: The implementation of quantitative models of real phenomena that are evolved on computational devices has become a common practice in science and engineering in the last 50 years. The major advantage of this technology is the possibility to process large amounts of data in relatively short times. In this review the major techniques of computational molecular science as applied in the industry of life sciences are reviewed. Further these techniques are discussed in view of their documented impact on the research and development workflow, with real illustrative examples taken from the nutrition and food industry. Computational molecular sciences have now come of age and, when deployed knowledgably, are shown to be able to generate intelligent hypotheses for project advancement, to lower attrition rate and to ultimately reduce research and development costs and to shorten the time-to-market cycle. The documented successes are however episodic and therefore efforts should be directed towards the development of standard reproducible protocols to use this technology.

Keywords: Cheminformatics · Computational molecular science · *in silico* Screening · Molecular docking · Molecular similarity

Introduction

In the last 50 years the application of computers and computer programs in everyday life and more specifically in science has developed very rapidly. Beside the obvious simplification of repetitive tasks, computers can solve complex mathematical models of real objects and thus bring into reach what was formally the privilege of a few intuitive and visionary minds. The trend is well illustrated when considering the application of computational technology to the development of biologically active compounds. Initially, discovery was based on the classical pharmacological approach, but during the Renaissance reductionist approaches started to dominate in the natural sciences. Only one or two active components were extracted from plant preparations endowed with biological activity. Almost at the same time, atomistic theory came back to the scientific 'goût du jour' after a thousand-years-long absence. The molecular theory of matter became broadly recognized in the scientific community, after a series of successful validation experiments. In the early 20th century, the Schrödinger equation was proposed to relate molecular structure with physical properties. However, the attempts to relate molecular science to biology turned in a complex puzzle that we start to understand only today with the extensive help of data and models processed by computational devices.

This review contains three major parts: 1) The basic state-of-the-art techniques of computational molecular science; 2) Documented evidence that this young scientific discipline now is coming of age and is able to significantly impact the industry of life sciences; 3) Personal experience in designing and implementing a computational molecular science platform to support next generation nutritional research at Nestlé Research Centre. This effort is illustrated with examples of successful applications. In a conclusion the emerging trends in the post-genomic design of bioactive molecules is briefly outlined.

*Correspondence: Dr. M.G. Grigorov
Bioinformatics Group Leader
Nestlé Research Center
BioAnalytical Science Department
CH-1000 Lausanne 26
Tel: +41 21 785 8939
Fax: +41 21 785 9486
E-Mail: martin.grigorov@rdls.nestle.com
#This contribution is dedicated to my father Dipl. Ing. Chemist Grigor V. Grigorov on the occasion of his 65th birthday

The Techniques of Computational Molecular Science

Computer Hardware and Software

Access to visualization workstations and to a high-performance computing system is a condition *sine qua non* for the efficient implementation of a computational molecular science platform. With the advent of open-source software and the free Linux operating system, the costs of high-performance computing have decreased significantly. Most of the software providers nowadays distribute their products ported to the Linux operating system and a number of open-source modeling packages are available as alternatives.

Ligands

To undertake any investigation on large collections of molecules these should first be represented in a mathematical metric space. A variety of computable molecular properties are available [1] which can be divided in three major groups: 1D (one-dimensional) global descriptors; 2D (two-dimensional) topological descriptors, and finally spatial or 3D (three-dimensional) descriptors that are designed to capture the shape and surface properties, expected to have an important contribution to the recognition of ligands by macromolecules. These descriptors are often referred to as pharmacophores. Several validation studies indicated that similarity selection and design of virtual libraries is efficient enough when

using only 1D and 2D descriptors [2]. The similarity principle formulated by Johnson and Maggiora [3] reflected the experience of researchers in the area of medicinal chemistry, and was later elaborated by Frye [4] in the structure–activity relationship homology approach. The currently used metrics are based on the Euclidean distance and cosine angle functions in the case of real-valued descriptors, and the Hamming and Tanimoto distances when considering structural bit string descriptors [5]. An ideal similarity metrics should correlate with bioactivity in order to reflect the molecular recognition process in a realistic way. However, with the current lack of a set of such 'universal' descriptors, probably the best strategy is to use the most diverse molecular coding possible to capture the different structural and functional features of the investigated molecules. The more descriptors that are used to represent the molecular entities, the greater the likelihood that some will be correlated. To remedy such problems a number of dimensionality reduction algorithms are usually applied [6][7]. In the final stages of a discovery project computational bioavailability and computational toxicological assessments of the ligands are applied to lower the attrition rate in clinical trials. In most of the cases the computational evaluation of bioavailability is based on the Lipinski rule of five [8], but recently more sophisticated procedures to predict 'drug likeness' were published [9]. Computational toxicology is an emerging discipline aimed at raising early alerts for possible toxic issues, using quantitative structure–toxicity relationships [10].

Receptors

The sequencing of the human genome generated a large number of candidate protein targets. The evidence that the number of possible protein structures is rather limited compared with the huge number of functional sequences provided the drive for the launch of experimental structural genomics initiatives. This resulted in the public availability of the structure of some 15'000 proteins [11]. Concurrently, computational methods were described to provide in some cases faster and cheaper ways to generate protein structures based on homology modeling and fold recognition [12]. Such methods put now at reach the structures of a large number of pharmaceutically relevant protein targets.

At the same time, structure-based molecular docking techniques were developed with the goal of understanding in a better way biological function that originates in the molecular interaction of proteins with specific substrates [13]. The methodology permits the prediction of the binding mode of a molecule in a target protein and the estimation of the binding energy of the resulting interactions. Molecular docking consists of a conformational sampling part and an interaction energy scoring part. The sampling algorithm generates different binding modes, also called poses, of the ligand within the receptor in order to predict those that can stabilize the formed complex. Scoring functions evaluate the placement and the orientation of a given conformation regarding its physico-chemical and geometrical complementarity with the relevant receptor. Currently, major efforts are directed in improving the efficiency of the docking algorithms and the sensitivity of the scoring functions to discriminate the biologically relevant binding modes. Moreover, the methodology is evolving from the classical lock-and-key and induced-fit technologies to integrate the flexibility of the binding-site residues and backbone, and even of the whole macromolecular target [14]. To compensate for the inaccuracy of structural data due to poor resolution, or for conformational changes upon complex formation, low-resolution docking techniques have been developed that were able to predict the rough structural features of several ligand–receptor complexes, based on geometric complementarity [15]. In the special case of investigating substrate–enzyme recognition, standard molecular docking techniques could not provide any information about the possible chemical conversion of the substrate, catalyzed by the enzyme with, quite often, the help of a metal ion. The state-of-the-art technique used in such cases investigates the enzyme active site by quantum mechanical methods, while simulating the surrounding protein by molecular mechanics. The combined method is often referred to as the quantum mechanics/molecular mechanics (QM/MM) method, and is implemented in most software packages [16].

Material Science

Computational material science is aimed at understanding the properties of materials and at simulating their behavior at atomic, nano, microscopic and macroscopic levels. In this scale of about eight orders of magnitude, some properties of matter can be inferred from first principles using periodic boundary conditions that allow for the simulation of large systems by treating a small number of atoms. In other cases coarse-grained approaches allow simulations on plastic deformation using multi-millions of atoms. Hybrid methods were proposed which treat the atomistic details affecting the behavior of a material in a local region and combine it with the continuum description of the rest of the system. Computational material science simulations typically use molecular dynamics to simulate the time evolution of systems, whereas a spectrum of potentials is applied to model the physical interactions among the particles, ranging from *ab initio* quantum mechanical ones to coarse-grained interactions among beads. Computer simulations are becoming a very perspective tool in material science, as the approach is able to provide detailed information that is not obtainable from experiments.

The Impact of Computational Molecular Science on the Life Sciences Industry

When deployed knowledgably, computational molecular science has the potential to play an important and diversified role in the contemporary life sciences industry due to its capacity to generate intelligent hypotheses for project advancement, to lower the attrition rate, and to ultimately reduce research and development costs and time [17]. Typically the discovery pipeline of a life science company spans seven major phases. The first phases are target identification and target validation, followed by lead identification and lead optimization. If the successful leads survive preclinical and clinical evaluations they could be marketed as products [18]. In 2001, the Boston Consulting Group carried out a survey study that encompassed some 50 life science companies and academic institutions [19]. It turned out that the development of a new chemical entity (NCE) following classical approaches costs on average $880 million, taking about twelve years from target identification to regulatory approval. Of this cost, about 75% was attributed to failures along the product pipeline. The Boston Consulting Group estimated that computational molecular science, or *in silico* chemistry as they labeled it, alone has the potential to generate savings of up to $130 million, and to shorten the time-to-market cycle by almost a year. Two main approaches are the most relevant for this, namely chemical-similarity searching over large molecular databases and structure-based drug design (SBDD) of smaller compound collections [17]. As an example of the successful application of 3D-pharmacophore searching techniques, one might cite the discovery of a novel fairly potent dopamine transporter (DAT) inhibitor, while another report described the discovery of VLA-4 integrin antagonists with submicromolar potency [20]. As an alternative approach, 'structure-based' ligand design has matured into a viable method for the identification of hits and knowledgably contributed to the development of marketed drugs, such as Viracept sold by Agouron, or Relenza, marketed by Glaxo SmithKline. It is estimated that up to now, SBDD overall contributions amounted to the introduction of nearly 50 compounds

into clinical trials and to numerous drug approvals [21].

The application of computational molecular science is demonstrably able to generate genuinely novel leads at much higher hit rates compared to classical empirical high-throughput screening and substrate-based designs applied by medicinal chemists. But probably because a lead is still a long way from a marketable product, the souvenir of an original and valuable contribution of the technology is often lost. Therefore the most disturbing question that a computational chemist might still have to face is "Is there a case where a drug that is on the market was truly designed by a computer?" [17]. If the answer to this is 'no', the question itself is certainly not the most appropriate way to benchmark a technology. In fact there is not a single technology currently applied in the industry that can claim such a success. Rather, the correct way to judge the usefulness of a technology is to verify three more specific points, *i.e.* 1) to what extent the technology is able to bring competitive advantage to the company using it; 2) to what extent would the technology enable the company to cut down R&D costs and to shorten the time-to-market cycle; and most importantly; 3) to what extent is the performance of the technology sustainable and reproducible. Computational molecular science has had numerous successes in the first two categories, but it should be admitted that these remain episodic. The performance of the technology is not sustained and even expert practitioners are frequently surprised and disappointed about the variability of the achieved hit rates. Therefore efforts should be directed to the development of standard and validated protocols [17].

Computational Molecular Science Platform for the Nutrition and Food Industry

The implementation of computational activities in the major areas of the biotechnology industry should be conducted according to the specificity of each of the businesses. In the following I will briefly review how this was achieved at Nestlé Research Centre by recognizing the specificities of the food and nutrition business (Fig.).

In the past much of the efforts in the life science industries were focused on the development of therapeutics for recovery from different diseases, but today the role of food and nutrition is increasingly recognized to be important in preserving the healthy life of the consumer as well as to counteract the onset of illness. While pharmaceutical intervention is making use of highly active synthetic molecular agents to reach specific pharmacological targets, nutritional preparations are made of exclusively natural ingredients aimed at providing a mild, long-term homeostatic effect. This is probably a first specificity of the nutritional industry. In order to respond to the challenge to identify novel bioactive ingredients we apply cheminformatics tools for the maintenance of a large virtual database containing some 110'000 natural compounds. To allow for chemical similarity searches in the database the compounds were encoded as 2D and 3D structures by using both commercial and in-house developed descriptors. The database is routinely screened for lead discovery in the early phase of every project entering the laboratory.

A second specificity of the nutrition and food industry is that it has a significant advantage over other life science businesses in its knowledge and application of the complex relationships established between the consumer and its symbiotic gut microflora, referred to as probiotic bacteria [22]. The typical probiotic bacteria belong to two distinct strains *Lactobacillus* and *Bifidobacterium*. To facilitate the proliferation of probiotic bacteria, nutritional research identified prebiotic fibers made of natural polysaccharides that are used as nutrients exclusively by the probiotic strains. Interestingly, the investigations carried in the laboratory indicated that the prebiotic fibers and their fragments might not only promote the growth of probiotics but also physically inhibit the attachment of pathogenic bacteria to the gut epithelial cells. This observation was made while studying the structural mechanism of enteropathogenic *Escherichia coli* (EPEC) attachment to gut epithelium. The host–pathogen interaction requires an intimate adhesion mediated by the bacterial adhesin intimin and another bacterial protein, the translocated intimin receptor (Tir) that is injected by the bacteria in the host cell and redisplayed at its surface. At the time when the project was initiated, a crystallographic structure of the intimin-Tir complex became publicly available [23]. We therefore applied structure-based virtual screening with a subset of some 8'000 different oligosaccharides and glycoconjugates from our database to identify a few leads tightly binding to the intimin protein, that were thus suspected to block the interaction with its counterpart Tir. This hypothesis was further confirmed by *in vitro* binding experiments of intimin and Tir with and without pre-treatment with the most potent of the leads. Interestingly, this most potent lead was found to be a glycohydrolytic product of a major component of cereal cell walls that is abundant in cereal preparations such as dough.

A third major specificity of the nutrition and food industry is the care with which food structures are designed at the molecular level. Proteins, carbohydrates and fats, together with air and water, are the main building blocks of food structures. The various three-dimensional combinations of these give rise to different food textures, like the crispness of fresh baked wafer, the smoothness of an ice cream, the sponginess of bread. In fact, the structure largely defines the food product, including its appearance, its texture, its solubility, as well as the release of its aromas and flavors in the mouth and the nose cavity and the availability of

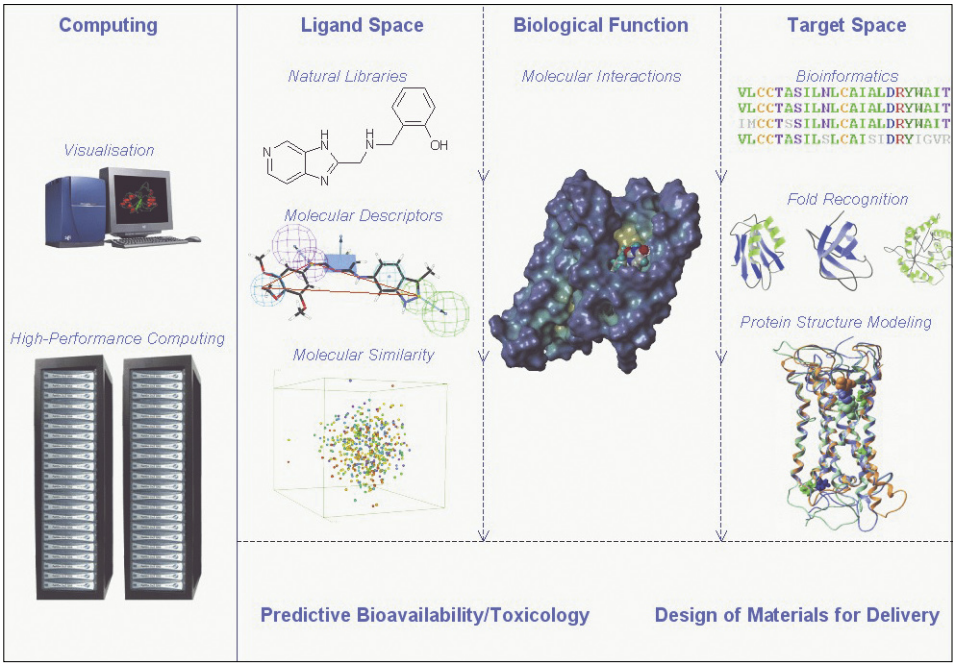

Fig. The basic competences of computational molecular science and the typical workflow in which they are organised

its nutrients in the digestive system. The application of computational material science was able to provide interesting insights in the search for ways to improve the solubility of cocoa powders. Powder particles were simulated as surfaces with different hydrophobicities, where the totally hydrophobic surface was modeled as an ordered array of alkanethiol chains attached to gold atoms. Perfectly spherical water droplets were consequently positioned on the model surface and the system was evolved by constant temperature and pressure molecular dynamics for several nanoseconds to predict contact angles upon relaxation of the liquid on the surface in full agreement with the experimental measurements.

An important specificity of nutrition and food industry is the recognition that food choice is influenced by an individual memory of food-related smells and tastes, built from the earliest moments of life. A final product is therefore never successful unless it tastes and smells good. Taste is categorized in five basic categories: salty, sour, sweet, bitter and umami, the later being characterized by the savory taste of monosodium glutamate (MSG). In a recent project we employed molecular-similarity-based virtual screening to identify novel umami-tasting compounds as alternatives to MSG. The hit compound monosodium N-acetylglycine was found to elicit umami taste at a 4-fold higher taste threshold than MSG, and was food-grade and cheap to produce [24].

Finally, the most important specificity of the nutrition and food industry remains its ability to respond to the consumer's expectation for safe foods. Major efforts were deployed to carefully scrutinize every single product, and as a result the classical risk of bacterial contamination is no longer an issue. However potential hazards emerge from environmental contamination of food raw materials such as from car exhausts or agricultural chemicals and their metabolites, for example. We contributed significantly to the development of the competency of computational toxicology that was integrated within an early warning system to assess the potential toxic hazard of food contaminants. For this the classical methods of cheminformatics and quantitative structure–activity relationships (QSAR) were applied.

Conclusions

A major trend in the contemporary life sciences industry is emerging due to the decline of the 'one-disease-one-target-one-ligand' approach that dominated thinking for a century [25]. It was realized that in biological systems everything is connected to everything else and recently major achievements were reported in our understanding of the global organization of cellular networks [26]. Computational molecular science has come of age and is entering an exciting new phase where the biological targets will shift from single proteins, to functional protein complexes, to whole networks determining precise cellular states, and where new foods and drugs will no longer be made of single active molecules but will represent molecular cocktails or multiple ligands [27] with components targeting the neural centers of whole disease-associated molecular networks. Nanotechnology offers the promise of enabling the health beneficial effects of such molecular cocktails through their controlled release by food raw materials transformed in delivery vectors.

Acknowledgments

The author would like to thank wholeheartedly Professor Dr. Jacques Weber for having supervised his Ph. D. work at the University of Geneva and for continuous support and encouragement. Thanks are also addressed to colleagues Dr. Ziding Zhang and Dr. Laurence Miguet who contributed to many aspects of the presented work.

Received: May 19, 2005

[1] R. Todeschini, V. Consonni, in 'Handbook of Molecular Descriptors – Methods and Principles in Medicinal Chemistry Series', Eds. R. Mannold, H. Kubinyi, H. Timmerman, Wiley-VCH, Weinheim, New York, **2000**.

[2] R.D. Brown, Y.C. Martin, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.

[3] M.A. Johnson, G.M. Maggiora, in 'Concepts and Applications of Molecular Similarity', Eds. M.A. Johnson, G.M. Maggiora, J. Wiley and Sons, New-York, **1990**, pp. 1–13.

[4] S.V. Frye, *Chem. Biol.* **1999**, *6*, R3–R7.

[5] P. Willett, J.M. Barnard, G.M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[6] W. Cooley, P. Lohnes, in 'Multivariate data analysis', Wiley, New York, **1971**.

[7] D.K. Agrafiotis, V.S. Lobanov, F.R. Salemme, *Nat. Rev. Drug. Discov.* **2002**, *1*, 337–346.

[8] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeny, *Adv. Drug. Deliv. Rev.* **1997**, *23*, 3–25.

[9] a) A. Ajay, W.P. Walters, M.A. Murcko, *J. Med. Chem.* **1998**, *41*, 3314–3324; b) J. Sadowski, H. Kubinyi, *J. Med. Chem.* **1998**, *41*, 3325–3329; c) A. Wagener, V.J. van Geerestein, *J. Chem. Inf. Comp. Sci.* **2000**, *40*, 280–292.

[10] A.H. Hall, *Toxicology Letters* **1998**, *102–103*, 623–626.

[11] J.C. Norvell, A.Z. Machalek, *Nat. Struct. Biol.* **2000**, *7*, S. 931.

[12] K. Mizuguchi, *Drug Discov. Today.: Targets* **2004**, *3(1)*, 18–23.

[13] P.D. Lyne, *Drug Discov. Today* **2002**, *7(20)*, 1047–1055.

[14] H.A. Carlson, *Curr. Opin. Chem. Biol.* **2002**, *6(4)*, 447–452.

[15] E. Katchalski-Katzir, I. Shariv, M. Eisentstein, A.A. Friesem, C. Aflalo, I.A. Vakser, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 2195–2199.

[16] A. Warshel, *Ann. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425–43.

[17] a) K.H. Bleicher, H.-J. Böhm, K. Müller, A.I. Alanine, *Nat. Rev. Drug. Discov*. **2003**, *2*, 369–378; b) W.L. Jorgensen, *Science* **2004**, *303*, 1813–1818; c) X. Barril, R.E. Hubbard, S.D. Morley, *Mini-Rev. Med. Chem.* **2004**, *4*, 779–791; d) B.K. Soichet, *Nature* **2004**, *432*, 862–865.

[18] J.-B. Léauté, *Chin. J. Chem.* **2003**, *21*, 1241–1246.

[19] P. Tollman, P.A. Guy, in 'A Revolution in R&D, How Genomics and Genetics are Transforming the Biopharmaceutical Industry', Andersen Consulting, BCG Report, **2001**.

[20] a) S. Wang, S. Sakamuri, I.J. Enyedy, A.P. Kozikowski, O. Deschaux, B.C. Bandyopadhyay, S.R. Tella, W.A. Zaman, K.M. Johnson, *J. Med. Chem.* **2000**, *43*, 351–360; b) J. Singh, *J. Med. Chem.* **2002**, *45*, 2988–2993.

[21] L.W. Hardy, A. Malikayil, *Curr. Drug Discov.* **2003**, 1–6.

[22] A.T. Borchers, C.L. Keen, M.E. Gershwin, *Handbook Nutr. Immun.* 2004, 213–241.

[23] Y. Luo, E.A. Frey, R.A. Pfuetzner, A.L. Creagh, D.G. Knoechel, C.A. Haynes, B.B. Finlay, N.C. Strynadka, *Nature* **2000**, *405*, 1073–1077.

[24] M.G. Grigorov, H. Schlichtherle-Cerny, M. Affolter, S. Kochhar, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1248–1258; M.G. Grigorov, H. Schlichtherle-Cerny, M. Affolter, S. Kochhar, M.-A. Juillerat, 'Flavoring compositions containing N-acetylglycine as umami taste enhancer', EP 1356744 A1 **2003**.

[25] J.R. Sharom, D.S. Bellows, M. Tyers, *Curr. Opin. Chem. Biol.* **2004**, *8(1)*, 81–90.

[26] M. Grigorov, *Drug Discov. Today* **2005**, *10* (*5*), 365–372.

[27] R. Morphy, C. Kay, Z. Rankovic, *Drug Discov. Today* **2004**, *9*, 641–651.