Inverse Problems and Hierarchical Multiscale Modelling of Biological Matter

Alexander Lyubartsev and Aatto Laaksonen

Abstract In solving mathematical and physical problems we generally think that the problem can be condensed to a well-defined equation which then can be solved either analytically or numerically with the provided input data and, if necessary, applying initial and boundary conditions to limit the amount of solutions to something which makes sense and can be accepted as the correct solution. This is the typical practice in "forward" problem solving. Consequently the *solution* can be considered as the *inverse* of the problem. Indeed, it is not uncommon in Science and also in everyday life to have the "solution" without knowing what exactly did "cause" it and "how". We take an illustrative example from Forensics with a case where a lethal crime is committed with a dead body (solution), but at the time of arrival to the crime scene the details to start to investigate are scattered all over the place. The investigators have the difficult task to mentally reverse the time to get good enough picture of the crime (problem) to start to trace the criminal and murder weapon (input) and possibly also the motive (cause). In this Chapter we discuss on how solving of the Inverse Problems is entering in Chemistry and focus on our own inverse computer modelling method to create a model (force field) from the results we already have. We explain how this method, called the Inverse Monte Carlo, can also be used for systematic hierarchical multi-scale modelling based on successive coarse-graining from first-principles to meso-scale and even further by super-coarse-graining. We show several applications of using it and also vision future prospects of hierarchical multi-scale modelling.

A. Lyubartsev · A. Laaksonen (✉)
Department of Materials and Environmental Chemistry, Arrhenius Laboratory, Stockholm University, 106 91 Stockholm, Sweden
e-mail: aatto@mmk.su.se

A. Laaksonen
Centre of Advanced Research in Bionanoconjugates and Biopolymers, Petru Poni Institute of Macromolecular Chemistry, Aleea Grigore Ghica-Voda, 41A, 700487 Iasi, Romania

State Key Laboratory of Materials-Oriented and Chemical Engineering, Nanjing Tech University, Nanjing 210009, People's Republic of China

Department of Engineering Sciences and Mathematics, Division of Energy Science, Luleå University of Technology, 97187 Luleå, Sweden

M. J. M. Abadie et al. (eds.), *New Trends in Macromolecular and Supramolecular Chemistry for Biological Applications*, https://doi.org/10.1007/978-3-030-57456-7_11

1 Introduction

1.1 Inverse and Forward Problems

Solving the Inverse Problems and practising "Reverse Engineering" must have appeared naturally out of curiosity and necessity along with direct problem-solving we humans have been faced to, first for survival and conformity, successively leading to civilisation, technology and Sciences. Solving Inverse Problems in Science is first of all a mathematical problem (Tarantola 2005; Yaman2013), common in many areas from Chemistry and Biology to Medicine and Astronomy just to mention a few. In Medicine doctors regularly need to suggest a diagnosis and treatment to the illness of the patient, while in Astronomy we may have a mysterious spectrum travelled from a very distant star. In the first example the inverse problem can be solved more or less rigorously either by systematically recognising patterns or empirically and even by guessing, while in the second example there are scientific methods to solve it backward by analysing the wave lengths giving information about the source (and its chemical composition) as a cause of the distinct radiation. For many living species in Nature unconsciously solving of the inverse problem is critically important for their very existence in catching food or navigating.

To solve forward problems can be easy and quickly done "on the back of an envelope" but it may also require powerful supercomputers and complex numerical algorithms to provide the correct result. Modelling a multimillion ensemble of strongly interacting particles to describe biological systems by solving Newton's equations is a typical example of the latter. Solving inverse problems can become often mathematically incredibly difficult simply because they tend to be improperly posed so that several inputs can give same results as a cause of the forward problem even when the data appears perfectly correct. For example two (or more) different sound sources can produce the same overlapping acoustic pattern. Therefore a unique correspondence is an important requirement for a complete solution of the inverse problem.

2 Design in Chemistry

2.1 Forward Design

In Chemistry and in particular in Materials Science, one particular word has lately come very popular and it is "Design". We now can access and manipulate the building blocks of matter using a variety of experimental techniques and since all matter around us and also inside us is made of molecules, we should be in a good position to "design" material with specific properties and function. However, since there is roughly ten orders of magnitude difference between molecular sizes and sizes of

macroscopic stock material to manufacture products for us humans this design is not always easy or straight-forward. Would it be easy straight-backward? There are many ideas now emerging around it,

Solving inverse problems has during the last decades become an important instrument in many scientific disciplines. In particular they have recently gained popularity in Chemistry and in Materials Sciences. A recent development is called "Inverse Design" (Sanchez-Lengeling and Aspuru-Guzik 2018). However, before introducing it let us first discuss the "Forward Design" which has been the common practise since early attempts to gain metals create explosives and process food and necessities. Finally stoichiometry allowed quantitative studies and chemical synthesis was discovered. Chemistry and chemical engineering are based on and develops with the accumulated and documented knowledge. Chemical intuition of its practiser is always an important component leading to new knowledge. There is no denying that the so called "trial-and error" and "cook-and-look" techniques have been common in many Chemistry laboratories to produce new chemicals and substances with specific desired properties and functions. Serendipity, the complete opposite to design, did appear occasionally to help to give the most important results. Thanks to computers continuously becoming more powerful and numerical search algorithms more efficient together with rapidly growing well-organized and easily accessible molecular and chemical databases, the "Forward Design" has radically changed its strategy from cooking molecules in laboratories to mining of existing molecular data using computers as an important component to shorten the cycle from ideas to products.

2.2 Knowledge-Based Design in Chemistry

During many decades now the chemical research has produced an extensive amount of new molecules with their specific properties tagged, organized and stored in large data bases. This data comes both from experiments and theoretical studies. Having this accumulated knowledge and data publicly open and easily accessible, as well as its correctness verified before storing it, can prevent us making (same) mistakes or discovering the wheel again, but most importantly to help us to discover new knowledge and correlations still hidden there. Along with these big data mines, new disciplines are developed of how to extract the desired information from it (Informatics, Chemometrics, Machine Learning, etc.), we have new options and methods to produce novel materials. By starting to search specific type molecules in the "Molecular Space" containing now virtually all possible molecular structures and by screening to gradually select a "Chemical Space" containing only a handful of candidates having the chemical properties we are looking for. Pharmaceutical industry has many decades been using this type of rational design based on virtual screening and quantitative structure-activity relationship (QSAR) models with carefully designed descriptors to find new compounds (leads) to become efficient medicines with optimal therapeutic effects while showing a minimum amount of side effects. With machine learning techniques entering Chemistry more and more simple regression models are

being substituted with more intelligent methods such as neural networks and others to find hidden correlations in big data structures and even new "physical" rules.

2.3 Inverse Design in Chemistry

There are many different strategies to solve inverse problems in Chemistry. Typically in utilizing the "Inverse Design" we can do the opposite to "Forward Design" and start by specifying the property or function of the new material we then look for from the "Chemical Space" and thereafter start to screen the content in "Molecular Space" with suitable descriptors trying to successively narrow the space to a small number of molecules having the property and function we look for (Sanchez-Lengeling and Aspuru-Guzik 2018). This type of search is strongly aided with machine learning and other techniques of artificial intelligence now increasingly applied in Chemistry. The types of materials inversely designed this way so far contain mainly drugs and organic and inorganic materials for optical and electronic devices, including batteries (Sanchez-Lengeling and Aspuru-Guzik 2018). Other approaches to design specific molecules also include so called Variational Particle Number and Alchemical Potentials within density functional theory (von Lilienfeld et al. 2005), Linear Combination of Atomic Potentials (LCAP) (Wang et al. 2006), which can be implemented in classical and quantum mechanical (DFT, tight-binding DFT and time-dependent DFT) Hamiltonians to perform many types of property optimizations. Inverse band-structure problem is solved for finding an atomic configuration with given electronic and optical properties (Franceschetti and Zunger 1999). Struebing et al. (2013) suggest an inverse method to find an optimal solvent to maximally speed up the kinetics for chemical reactions. Inverse methods can be used successfully in Spectroscopy for example to decompose 2D NMR spectra of mixtures of molecules utilizing the technique of blind source separation (BSS) (Cherni et al. 2019). Jonas uses a so called deep imitation learning protocol to solve the chemical structure from molecular formula and NMR spectrum (Jonas 2019). There are many excellent reviews on these topics,such as Machine Learning which is strongly entering into Computational Chemistry and Materials Science via knowledge-based modelling is excellently reviewed in Ferguson (2018), Butler et al. (2018). Several emergent methods of Inverse design are reviewed in (Sanchez-Lengeling and Aspuru-Guzik 2018; Noh et al. 2020; Sherman et al. 2020; Hu et al. 2009; Martinez-Luaces 2012; Hachmann et al. 2018).

2.4 Solving Inverse Problems in Computer Modelling and Simulations

Ideas of solving the inverse problems in Computational Chemistry have their roots in liquid state theories of mid last century (Kunkin and Frisch 1969) and started gain popularity in the following decades along with the development of more powerful computers (Rosenfeld and Kahl 1997). In computer simulations the most important input is the interaction potential or Force Field. It can be constructed conceptually based on simple models used to describe interatomic or intermolecular interactions, all having their fundamental origin in how electrons and nuclei in atomic and molecular frameworks feel the presence of each other and the forces they give rise to. Long before the computers did arrive scientists knew that there was a weak attraction even between neutral particles (atoms and molecules) at long distances (order of a few Ångström) which grew stronger at closer distances and became strongly repulsive at short distances. Lennard-Jones (Jones 1924) and contemporary scientists proposed very simple mathematical models to describe it. For charged particles there was Coulomb law describing their mutual strong and long-range interactions. The assumption of additivity of the interactions and that only two particles did interact with each other momentarily were assumed and simplified the theoretical work as effective pair potentials became a common tool to model interacting particles. When parameterized on experimental information they could contain a certain part of many-particle character built in thereby improving their quality.

When studying condensed phases of matter the structural information can be obtained from the pairwise correlations of the neighbouring particles (atoms, ions, molecules etc.). These correlations can be constructed statistically by sitting physically on each particle at a time and looking for the neighbouring particles in any direction radially and adding every hit in a histogram as a function of the mutual distances. When normalized to bulk density these pairwise correlations become radial distribution functions (RDF) or pair correlation functions giving probabilities to find particles at specific distances. Very early it was observed by scientists working in developing liquid theories and so called integral equation models that there was a correlation between pair correlation functions and pair potentials, Indeed, there is an explicit expression allowing, in principle, to compute the RDF from known pair potentials, and approximately it is always possible by particle-based computer simulations. An inverse problem, that is determination of pair potentials if RDF is known is however not a trivial task. Johnson and March did calculate potential of mean forces (PMF) from RDFs obtained in early diffraction studies of liquid metals and iterated them to pair potentials (Johnson and March 1963). An important foundation was the Uniqueness theorem of Henderson for RDFs and pair potentials for systems in equilibrium Henderson (Henderson 1974) which was later shown to be true even for multi-component systems. These theorems however did not say how to obtain one from the other. In 1979 Swendsen published a Monte Carlo method for statistical mechanical simulations and renormalization-group analysis of critical properties applied on the Ising model presenting an effective Hamiltonian and iterative solution

of the problem to obtain an effective pair potential from nearest-neighbor pairs of the spins (Swendsen 1979). Inspired by the work of Swendsen, in 1995 Lyubartsev and Laaksonen suggested a generalized inverse scheme to invert RDFs to effective pair potentials for arbitrary molecular systems (Lyubartsev and Laaksonen 1995), later known to be the Inverse Monte Carlo (IMC) or Newtonian Inversion (NI). This method is discussed in more detail in the next Chapter. Similar structure-based inversion methods have also been suggested (Soper 1996; Reith et al. 2003) being inspired by earlier works of Schommers (1983) and Reatto et al. (1986). Also the method of force-matching based inversion method of Izvekov and Voth should be mentioned (Izvekov and Voth 2005) which in turn was inspired by the force-matching method of Ercolessi and Adams (1994). In addition there are many other promising schemes including Generalized Yvon-Born-Green (YBG) method (Mullinax and Noid 2010; Cho and Chu 2009), relative entropy (Shell 2008) and configurational temperature (Mechelke and Habeck 2013).

3 The Inverse Monte Carlo Method

3.1 Theoretical Foundations

We will describe here the method of Inverse Monte Carlo (IMC). It is a mathematical method to coarse-grain (CG) the molecular interactions based on structural information obtained in underlying accurate studies (theoretical or experimental) of the same molecular system. Also equilibrium thermodynamic information can be used to obtain the CG model. It has some common features with many of the methods discussed above. The general idea in all coarse-graining is to produce a simpler model, or a kind of caricature, of the more accurate model full of details. Like in watching photos or paintings from a long distance the small details become less important. The same is true in modelling matter at longer length scales. Also at longer time scales the fast molecular fluctuations are averaged making molecules effectively more rigid in one or several conformations (geometries). This type of simplification is done in coarse-graining where we remove those degrees of freedom (DoF) from the original detailed potential energy Hamiltonian $H(r_1, \ldots r_\mathrm{n})$ which are less-important while we describe the system by keeping only the important DoFs $(R_1, R_2, \ldots R_N)$, where $\mathrm{N} \ll \mathrm{n}$.

How to choose the important DoFs is very much a matter of taste (or rather of experience based on chemical and physical intuition) as in many cases it turns out not to be very critical. As CG sites one often chooses centre-of-masses (COM) of the molecular group forming the CG. Generally, the CG coordinates are some functions of atomistic coordinates:

$$R_i = \theta_i(r_1, \ldots r_n) \tag{1}$$

Although in some cases the CG coordinates (R_i) may coincide with coordinates of certain atoms (r_j).

Once a mapping scheme of atomistic coordinates to CG coordinates (1) is chosen, we can describe a coarse-grained (CG) Hamiltonian:

$$H_{\mathrm{CG}}(R_1, ..., \mathrm{R}_N) = -\frac{1}{\beta}\ln \int \prod_{\mathrm{i}=1}^{n} \mathrm{dr}_i \prod_{\mathrm{j}=1}^{N} \delta\big(R_j - \theta_j(r_1, ..., \mathrm{r}_n)\big)\exp(-\beta \mathrm{H}(r_1, ...r_n)) \tag{2}$$

which is an effective N-body CG potential energy function. However, expression (2) cannot be used directly in CG simulations, therefore we first map or fit it to a pair potential:

$$H^{CG}(R_1, R_2, ..., R_N) \approx \sum_{i>j} U_{ij}\big(R_{ij}\big) \qquad R_{ij} = \big|R_i - R_j\big| \tag{3}$$

Many methods discussed in previous Section fit the Hamiltonian (3) to reproduce some properties from the underlying simulations with all the DoFs, for example energy, forces, or some features of structure. In IMC we use ensemble averages and RDFs as our first choice.

The IMC method is a powerful general method to invert ensemble averages, and particularly RDFs to effective pair potentials. It completely solves the inverse problem providing a unique solution. For any multi-component system it produces the effective pair potentials between selected sites (atoms, center-of-mass, and any other type of sites) in an inverse process which in forward process (input → model → simulation → results) reproduce completely the RDFs used as input and inverted to produce the effective potentials (model) when same conditions are applied. There may not be much point in doing the full circle this way but now we can increase the size of the system a few orders of magnitude and perform a simulation which we could not afford before when using the original detailed model, typically the all-atom (AA) model. The reason is not only that we have now much less interaction pairs already making simulations faster but the effective potentials between heavier sites are also softer and we can choose a longer time step. Still, the greatest effect using IMC comes from not needing to use any explicit solvent in the CG simulations. This will be discussed in more details below.

In the calculation of an RDF or g(r) in a typical particle simulation of N particles in a volume V the radial particle-particle distances are discretized to histograms S_α which after normalization to bulk particle density (N/V) show the probabilities to find the distances of the neighbouring sites between a chosen pair of atoms (or other type of sites). This histogram or the estimator of the pair-correlations is used in IMC. By discretizing the CG Hamiltonian (3) to two histograms of which one is the normalized estimator S_α from the g(r) and the other is an effective CG pair potential V_α in a table form we will obtain:

$$H = \sum U_\alpha S_\alpha \tag{4}$$

where ensemble average of S_α is related to RDF by:

$$g(r_\alpha) = \frac{1}{4\pi r_\alpha^2 \Delta r} \frac{V}{N^2/2} S_\alpha \tag{5}$$

Besides RDFs between non-bonded CG sites, other structural properties can be included into set of S_α, for example distribution of CG bond lengths, angles and torsion angles. This is very useful when we coarse-grain biomolecules (or other large molecules such as polymers), for which we create models consisting of CG sites connected by CG bonds. Bonded and angular potentials of such models are fitted to reproduce bond and angular distributions originated from the detailed model within the same IMC framework.

While in standard simulations (direct forward problem) we have interaction potential U_α as input, and we can evaluate $<S_\alpha>$ and RDF as output, in the coarse graining by IMC we solve the inverse task: from averages $<S_\alpha>$ (determined in atomistic simulations) we determine CG potentials. Solution of this non-linear inverse problem can be reached iteratively by the Newton-Raphson method, and by this reason the IMC approach is also referred as "Newton Inversion". Let us determine Jacobian of $S(U)$ dependence:

$$J = \frac{\partial \langle S_\alpha \rangle}{\partial U_\gamma} \tag{6}$$

as well as use the "vector" notations for the sets of potentials and RDFs:

$$\{\langle S_\alpha \rangle\} \equiv \vec{S}; \quad \{U_\alpha\} \equiv \vec{U}$$

Jacobian (6) expresses how changes of potential affect RDFs:

$$\Delta \vec{S} = \hat{J} \Delta \vec{U} \tag{7}$$

From the statistical-mechanical relationships this Jacobian can be computed by doing direct simulations with potential U:

$$\frac{\partial \langle S_\alpha \rangle}{\partial U_\gamma} = -\beta \left(\langle S_\alpha S_\gamma \rangle - \langle S_\alpha \rangle \langle S_\gamma \rangle \right) \tag{8}$$

Now we have everything to solve the inverse problem. We start simulation from a trial potential which can be the mean force potential $U_\alpha = -k_{BT} ln \langle S_\alpha \rangle$ or just zero. We compute from these simulation RDFs and $\langle S_\alpha \rangle$ and determine deviations from the reference RDFs $\langle S_\alpha^{ref} \rangle$ obtained in atomistic simulations:

$$\vec{\Delta} S = \vec{S} - \vec{S}^{ref} \tag{9}$$

Now we can compute which change in the potential is required to get the desired change in corresponding RDF:

$$\Delta\vec{U} = \hat{J}^{-1}\Delta\vec{S} \tag{10}$$

We update the interaction potential $\vec{U} \rightarrow \vec{U} + \Delta\vec{U}$ and repeat the procedure to compute a new corrected estimator. The iterative procedure is repeated until the estimators become identical (within statistical error of the simulations) to those in the original input g(r), also giving back the final RDFs obtained in the fine-grain simulation serving as reference. After convergence, we obtain the CG effective pair potentials which contain all the atomistic details to reproduce the microscopic structure found in the all-atom simulations, considering only the sites chosen to the CG model. It is like to visualize the AA simulation but leaving out the atomistic details and also the solvent. IMC fully includes the cross-correlations between the pair interactions, making the effective pair potentials more accurate in comparison with interactions obtained from using other similar methods. This is one of the properties making IMC a superior inverse method.

3.2 Using the Inverse Monte Carlo Method

In practical computations, and in particular for large molecules, direct use of expressions (8)–(10) may lead to non-convergence since the method is based on linear extrapolation (7) of a generally non-linear relationship. A simple way to regularize the procedure is to go by "small steps" to remain always in the linear regime and to multiply the change of RDF (9) (and respective change of potential (10)) by a damping factor $0 < \lambda < 1$.

What really makes IMC superior compared to other corresponding methods producing effective CG pair potentials is that even the solvent molecules can be considered as uninteresting DoFs. Therefore in using IMC CG pair potentials in simulations no explicit solvent is needed. The solvent is not implicitly there in the simulation cell as a continuum like in common implicit solvent models and most often characterized by the dielectric constant of the bulk solvent. The solvent is in the effective IMC CG potentials and contain specific atom-atom interactions for example to display H-bonds and solvation and hydration structures as in the underlying atomistic simulations. This is an extremely important feature as in all molecular simulations containing explicit solvent (most often water) it is the moving of the solvent around in the simulation cell which consumes major part of the computing time. In other words the IMC CG pair potentials are *solvent-mediated* containing the solvent in their functional forms. When used for example in simulations of biomolecular systems in water solution with ions and salt, all types of interaction forces are still there (hydrophobic, hydrophilic, H-bonds, steric, etc.). Self-assembly and other typical biological processes can be described accurately without explicit water molecules

as the water is still there in the effective potentials. The IMC potentials both supply the specific solute-solvent interactions and provide the collective effect of water to hydrophobicity and hydrophilicity as in AA simulations. Indeed, the IMC CG potentials for soft meso-scale particles while containing the atomistic interactions are so accurate that they can be coarse-grained again. This super-coarse graining allows us to perform reliable CG simulations of truly large systems like chromatin as will be described below.

The IMC CG pair potentials are produced in a tabulated form and their functional form can be very different from that of standard potentials like Lennard-Jones used often as empirically parameterized CG potentials. This is because it does have all types of intermolecular interactions together. Also, as it is solvent-mediated it has an oscillatory form at short distances reflecting the structure of solvation and hydration layers while this behaviour weakens and disappears at longer distances.

At long distances the IMC CG potential becomes overlapping with Coulombs law for charged particles as might be expected. Indeed, the Coulomb potential can be separated from the overall IMC CG potential and treated separately so that Ewald summation can be used in CG simulations. This divides in a natural way the CG potential to a specific short-range potential and long-range potential. Separating short-range and long-range Coulombic in CG simulations was also shown to produce more reliable overall results (Wang et al. 2013). Also, by making use of the inherent temperature dependency of the dielectric constant (Mirzoev and Lyubartsev 2011) the results obtained from IMC CG simulations can be made much less sensitive to temperature changes. In principle, the simulations with IMC CG potentials should be carried out at or close to the conditions applied in the underlying detailed simulations to calculate the RDFs as will be discussed below.

Intramolecular IMC CG potentials may also substantially deviate from the harmonic potentials bonded interactions are known to give rise to. However, in some cases they are close enough to harmonic so harmonic springs can be used, at least initially. In constructing IMC CG force field it is not necessary to follow the concept of dividing interactions for bond-stretching, angle-bending and rotation around bonds as in common atomistic force fields, as these can be substituted by a matrix of bead-bead interactions. IMC converges rapidly for systems with small molecules but for large biomolecular systems constructing converged potentials can become time-consuming and to do the inversion process manually becomes in practise out of question for biomolecular systems. Auxiliary tools are needed.

3.3 MagiC

There is a software package MagiC (Mirzoev and Lyubartsev 2013) developed which performs the structure-based coarse-graining of IMC described above for arbitrary molecular systems and models which has been recently updated to version 3 (Mirzoev et al. 2019). Magic is an essential tool and integral part of IMC to perform systematic coarse-graining. It can take the input from several common simulation packages

while it produces CG potentials to be run on many coarse-grained simulation program packages like the GPU-Accelerated LArge-scale MOlecular Simulation Toolkit (GALAMOST) (Zhu et al. 2013), the Large-scale Atomic/Molecular Massively Parallel Simulator LAMMPS (Plimpton 1995) and the GROningen MAchine for Chemical Simulations Gromacs (Van Der Spoel et al. 2005). MagiC software has many additional functions from diagnostics to visualisation.

3.4 Applications of Inverse Monte Carlo

The method of Inverse Monte Carlo was presented in mid-90's (Lyubartsev and Laaksonen 1995) and had to be soon applied on demanding systems to verify its power. 25 years ago a fairly "demanding" system was already to study NaCl in water even with the access to national supercomputers. We did study this solution at different salt concentrations. The number of molecules did range from 256 to 2000 with up to 20 ion pairs included. The simulations did cover a few nanoseconds (Lyubartsev and Laaksonen 1996). One reason why these were demanding simulations was that we did need very well converged RDFs for many different concentrations to accurately invert them to effective solvent-mediated Na-Cl Na-Na and Cl-Cl ion-ion potentials (Lyubartsev and Laaksonen 1997). We then did apply them to compute both the osmotic and activity coefficients for the hydrated ions. To obtain reliable results we did need 200 ion pairs. Systems of this size would not have been feasible to simulate using all-atom models with the computer power available for us at that moment. The water-mediated CG ion-ion potentials gave excellent results within the entire concentration range from 0.001 to 5 M. We did then perform a detailed MD simulation study about alkali metal ion (Li+, Na+, K+, and Cs+) condensation around double helix DNA in water solution (Lyubartsev and Laaksonen 1998) from which we obtained ion-DNA RDFs to invert them in IMC to effective CG potentials (Lyubartsev and Laaksonen 1999). Now already having water-mediated ion-ion CG potentials (we did compute them also for Li+, K+ and Cs+) we now had a complete set of CG potentials to perform simulations of a large double strand DNA chain in water in a box of $100 \times 100 \times 68$ Å^3. Using our water-mediated CG potentials we could find the order Cs+>Li+>Na+≈K+ for counter ion binding to DNA in agreement with several independent experimental studies.

We did also perform Car-Parrinello simulations for liquid water to compute effective water mediated CG potentials from first-principles MD simulations (Lyubartsev and Laaksonen 2000). In this way we calculated a new atomistic water interaction potential from RDFs obtained in simulations with elec-tronic degrees of freedom. We then did calculate hydration of Li+ ions from Car-Parrinello (Lyubartsev et al. 2001). In this way we could show that IMC was a true multi-scale modelling method where we could start from first-principles simulations and inverting the RDFs we would obtain AA interaction potentials and when using them in simulations (after increasing the size of the cell two orders of magnitude) we could construct a meso-scale effective pair potentials. This would allow us to hierarchically connect three

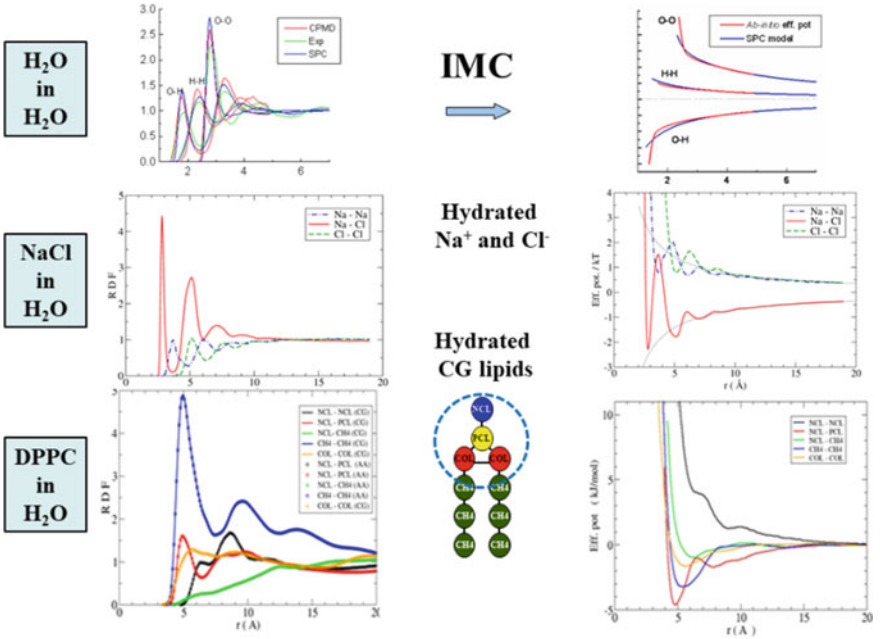

Fig. 1 RDFs from AA simulations and IMC inverted solvent-mediated CG potentials

scales: quantum, classical and mesoscale and start the whole multiscale modelling without any empirical information (Lyubartsev et al. 2009). We have previously summarized our work in (Lyubartsev et al. 2010; Lyubartsev et al. 2015) where there are more examples of the application of IMC. Much of the work done on biological systems is about nucleic acids.

In Fig. 1 we can see three different solvent-separated IMC CG effective pair potentials and the corresponding reference RDFs from which they are calculated. In all cases water is the solvent. Notice the oscillating nature of the potentials due to the presence of solvent built in them. In the top panel on the left we have RDFs from CPMD simulations of liquid water, classical simulations of SPC water and experimental water RDFs. On the right side are the corresponding inverted RDFs as pair potentials. For SPC we obtain back the SPC water model used in simulations to produce the RDFs. Note that although the RDFs are fairly similar the effective potentials from CPMD show much steeper repulsion compared to SPC model while at the long distance the models coincide as they should at Coulomb regime. Also the CPMD simulations of Li+ in water gave much steeper repulsion in the effective IMC ion water-oxygen potentials (not shown here, see Ref. (Lyubartsev et al. 2001) compared to common ion-water potentials models reflecting the exponential decay of repulsion based on Quantum Mechanical description. It could be fitted perfectly to $V_{\text{eff}}(\mathbf{r}) = A \exp(-B\mathbf{r})$ with $A = 37{,}380$ kJ/mol and $B = 3.63$ Å^{-1} indicating that Buckingham type of potential is to prefer. In fact, in the Lennard–Jones potential the

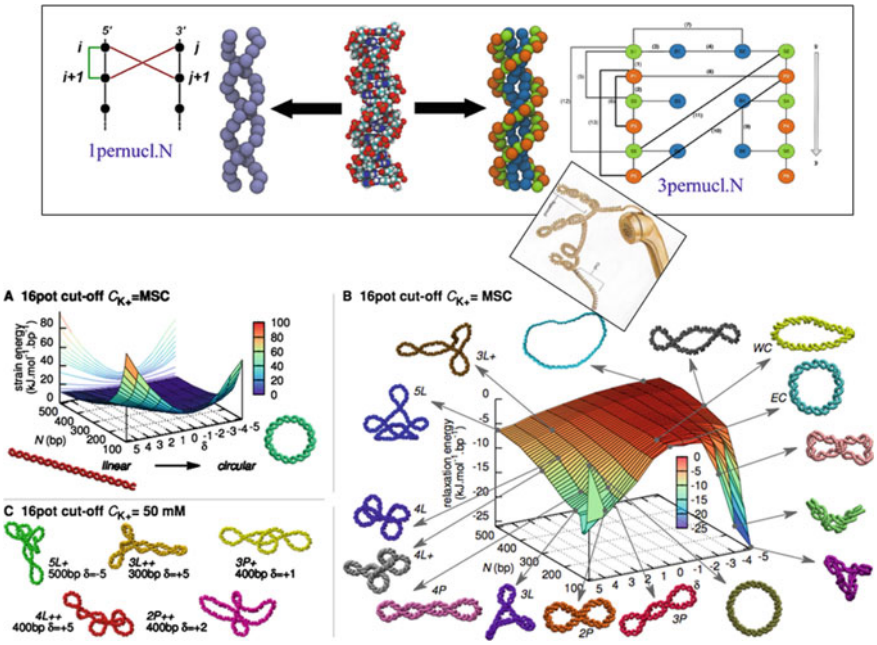

Fig. 2 Possible topologies (top) and twist and writhe oc minicircles of DNA (bottom) from Ref. (Naômé et al. 2014)

exponential decay of repulsion is approximated with the $1/r^{12}$ dependence making the calculations very much easier. In the middle panel are the water-mediated pair potentials calculated from the ion-ion RDFs, obtained in the aqueous solution of NaCl. Observe the distinct contact and water-separated potential wells in the Na^+ - Cl^- potentials and the additional hydration shells after them at longer distances seen as oscillations. At long distances they coincide with Coulombs law (dotted line) as can be seen in the figure. At the bottom there the IMC potentials for DPPC phospholipids obtained from AA simulations of randomly placed lipids in the water box. These IMC CG lipid models reproduce accurately all phases and morphologies these amphiphilic systems can form in water depending on the applied conditions. For more details see references (Lyubartsev and Laaksonen 1997; Lyubartsev et al. 2009, 2010). In Fig. 2 we apply IMC on DNA mini-circles up to 500 base pairs. In top two typical topologies are shown to coarse-grain DNA. Mechanical energies for linear and circular DNA are shown in bottom. For details see Ref. (Naômé et al. 2014, 2015).

3.5 *Transferability of IMC CG Potentials*

An important issue with the IMC CG potentials, and for that matter with most other CG potentials based on underlying fine-grain simulations, is that they describe the system at a specific thermodynamic state and the CG potentials automatically inherit the same condition and therefore should be used accordingly. It means that the effective potentials cannot be used to cover large intervals in temperature, concentrations, densities etc. But can they be used to study the same type of systems exhibiting very different topologies at same conditions? In our studies of multi-scale modelling of human telomeric quadruplexes (Rebic et al. 2015, 2017) we did construct the IMC potentials from AA MD simulations for the topomer, a [3 + 1] hybrid with a 26-nucleobase sequence d[AAAGGG(TTAGGG)$_3$AA] with K+ counter ions stabilizing the Hoogsteen structure (PDB id 2HY9). We later used successfully the same IMC potentials to model another quadruplex topology (PDB id 1KF1) also known to form from the human telomeric DNA sequence d[AGGG(TTAGGG)3] differing from 2HY9 in its loop topology and its G-strand relative orientation. The results are encouraging suggesting a certain degree of transferability in simulating quadruplexes with different topologies. Since the IMC potentials are normally not fit to any simple mathematical functions any general transferability is out of question due to their inherent complexity and difficulties to construct combination rules (Fig. 3).

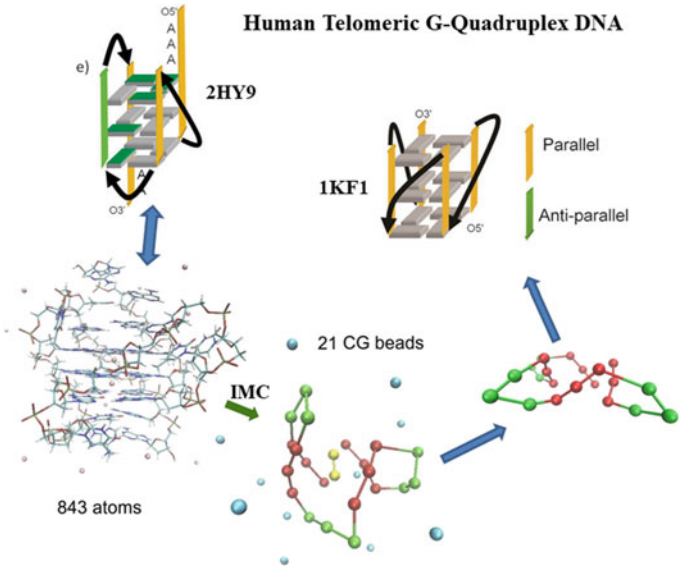

Fig. 3 IMC CG model for 2HY9 topology can be used in CG simulations of 1KF1. For more details see Refs. (Rebic et al. 2015, 2017)

3.6 IMC Effective Potentials and Dynamical Properties

An important issue in creating and using effective potentials based on structural information only is their capability to predict dynamical properties in CG simulations. It is reasonable to expect that CG potentials, being softer than AA potentials, create an energy surface where soft mesoscopic particles, even if heavier, move too fast compared to AA simulations or experiments. As no dynamical information goes in to the process in inverting the solely structural information to is also difficult to estimate the effect of its absence to dynamics of the particles. Also, even the fact that no explicit solvent is needed; making the simulations fast, the lack of friction normally generated by the solvent molecules affects the overall internal motion of the solute molecules. In our study of mini-circles of DNA (90–500 bp) at different salt concentrations using IMC CG potentials (Naômé et al. 2014, 2015), we did also study the dynamical properties for linear 18-mers of DNA in water solution with ions. We did observe that the counter-ions did diffuse roughly 220 times faster than in AA MD simulations while end-to-end fluctuations of DNA were 4.5 times faster and its total twist fluctuations 21 times faster than in AA MD simulations. Using Langevin dynamics gave us the opportunity to adjust the friction coefficient to give reasonable results for the dynamics. For example, using a friction coefficient 35 ps-1, which is comparable to the collision frequency of water molecules, brings the diffusion of the counter-ions close to the experimental values. In general as the degree of coarse-graining is different for DNA and for ions their dynamics differ so no general scaling factor is difficult to establish. Even if it is difficult to obtain quantitatively correct dynamics the relative trends can be reproduced normally well. Scaling CG potentials to reproduce experimental or AA-simulation results is needed otherwise.

3.7 Fine Graining

The opposite to coarse-graining is fine-graining (FG) or back-mapping, which is far from trivial as it requires inserting (re-inventing) the degrees-of-freedom which were deleted in coarse-graining. FG is not simple because the problem is under-determined. For polymeric systems, including biopolymers like nucleic acids, it is somewhat easier because the back-bone geometry and the topology of the monomeric units are known, this way providing several geometric constraints. For DNA there are the sequence information, the distance constraints for Watson-Crick base-pairs, the grooves and stacking to guide the FG. Common strategy is to start by a random placement, followed with a relaxation without adding any constraints. Thereafter check the chiralities, and first after that to start to apply the constraints. We have followed this strategy in (Lyubartsev et al. 2015; Naômé et al. 2014). Some software packages already have features for performing back-mapping, such as newer versions of Gromacs.

3.8 Bottom-Up Meets Top-Down

At the same time as multi-scale modelling based on the successive coarse-graining is a "from-bottom-up" technique to reach the mesoscale beyond, many sophisticated "from-top-down" experimental methods such as 3D imaging, microscopies, single-molecule manipulation and force measurements can come down to the meso-scale and even under. This means that these two approaches (bottom-up and top-down) can be made to meet at the mesoscale. A probable future scenario is mapping of images and forces from experiments on mesoscopic simulation models so that we start have good quality meso-scale force fields for CG simulations (as the AA force fields we have today) with experimental origin which also can be cross-fertilized with theoretical CG force fields inverted from underlying accurate FG simulations to make them more detailed. For example, AFM can supply vertical and lateral force-distance data which can be used to create a force field between particles and surfaces which is currently completely missing in modelling. All this would also allow us to improve the resolution of these experimental techniques all the way to the atomistic level by applying fine-graining techniques discussed in the previous section. Once atomistic level is reached quantum calculations can be performed to explore any possible reactive parts of the system. "Top-down meets bottom-up" will be a common theme in many integrative studies with multi-instrument experimental and multi-scale modelling is performed for the insight, discovery and design of novel products.

3.9 Advanced Use of IMC

We give here two more recent examples of multi-scale modelling based on hierarchical coarse-graining using the Inverse Monte Carlo method. The first deals with the important topic of bio-toxicity of nanomaterials and the second example is about super-coarse-graining of genetic materials.

3.10 Safety of Nanomaterials—Multiscale Modelling Using IMC

Nanomaterials ranging from carbon nanotubes and graphene structures to metal and metal oxide nanoparticles and quantum dots provide a virtually endless line of bioengineering applications, as well as offer use in modern nanomedicine as potential carriers for targeted drug delivery. On the other hand, the nanotoxic hazard associated with the penetration of small nanoparticles in biological tissues is a vividly debated subject. Concern is raised on both direct and potential long-term hazard to human health caused by nanomaterials since the immune system, developed during

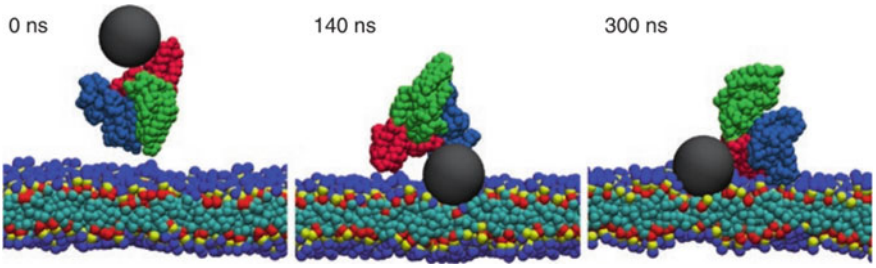

Fig. 4 Time sequence of CG simulation snapshots illustrating the interaction of a complex of negatively charged hydrophobic nanoparticle (2 nm radius) and Human Serum Albumin (HSA) protein with DMPC lipid bilayer. Reprinted from Hender et al., "Multiscale Modelling of Bionano Interface", in Modelling the Toxicity of Nanoparticles, L. Tran et al. (eds.), Advances in Experimental Medicine and Biology, Vol. 947, 2017, https://doi.org/10.1007/978-3-319-47754-1_7, with permission from Springer International Publishing

millions years of evolution, is not familiar with engineered nanoparticles. Existing methodologies of *in-vivo* toxicological evaluation require long time, are expensive and also connected with ethical concerns. This is why *in-silico* methods of prediction of toxic effects are of high demand (see Fig. 4). It is however extremely challenging task to model what is really happening from the first contact of a nanoparticle with biological matter on the molecular level to the effects of the nanomaterial on the whole organism.

Here we come to the point when hierarchical multiscale modelling, starting from atomistic description of the interface between nanomaterial surface and biomolecules (Brandt and Lyubartsev 2015) and proceeding to several levels of coarse-graining, handshakes with system biology, and in particular the Adverse Outcomes Pathways (AOP) concept (Halappanavar et al. 2020). An AOP relates the first biological effect of a nanoparticle taken up by an organism (called the Molecular Initiating Event), through a series of "Key Events" (measurable changes at organelle, cell, organ level), with an adverse outcome for the whole organism (and even for populations). First attempt to use multiscale simulations to predict some molecular initiating events in AOP, such as protein corona formations around a nanoparticle or membrane rupture using multiscale simulations have been recently published (Lopez et al. 2017; Power et al. 2019).

Necessity of validated tools for evaluation of potential toxic effects has led to appearance of "Safe by Design" concept (Schwarz-Plaschg et al. 2017) in design of nanomaterials. Toxicological safety should be considered as a compulsory property of all newly developed materials, alongside with the functional properties for which the material is developed. While performing the "Inverse Design", it is imperative to include the toxicological chain relating properties of molecular bio-nano interface with adverse outcomes, into the search in Chemical Space for specific materials satisfying these properties.

3.11 Super Coarse-Graining Using Inverse Monte Carlo

To apply hierarchical multi-scale methods there is hardly anything more suitable than the genome material which is strictly hierarchically built in consecutive orders (primary sequences, secondary, tertiary, quaternary, etc.) from underlying structures. The hierarchical order for modelling is: [DNA + histones] – [beads and strings of nucleosome core particles (NCP)] – [fibres of packed nucleosomes] – [chromatin loops] – [chromatin domains] – [extended and condensed chromosome]. Also, in terms of number of base-pairs there are six orders of magnitude from NCP to chromosome (10^2 to 10^8). We have in our previous studies demonstrated the importance of the electrostatic interactions and the power and accuracy of the IMC method by simulating these hierarchical building blocks in physiologically relevant conditions and with varied ion concentrations and observed the same condensation behaviour as in experiments while obtaining stable structures (Korolev et al. 2004; Korolev et al. 2006; Korolev et al. 2010). The CG simulation methodology developed by us is so detailed that even the coarse-grained model of nucleosome core particles can be coarse-grained while maintaining their molecular properties (Fig. 5).

We did simulate a clustering of 5000 NCPs obtaining an excellent agreement with small angle X-ray scattering spectra for a corresponding system (Fan et al. 2013; Sun et al. 2019). Also, we simulated large amount of DNA oligonucleotides (up to 400 oligonucleotides each of 100 base pairs of DNA) in presence of CoHex3+ ions

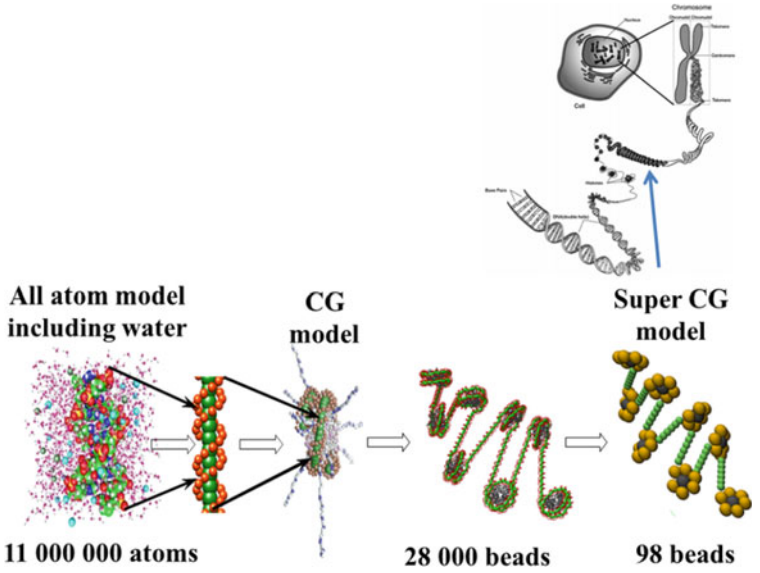

Fig. 5 Systematic hierarchical multi-sacle modeling from DNA to chromatin where higher order coarse-graining (super coarse-graining) is used. 98 beads corresponds a sysytem of 11 million atoms in corresponding all-atom simulation Ref. (Fan et al. 2010)

and found aggregation of DNA's in ordered hexagonal structures (Sun et al. 2019). Furthermore, using super coarse grained DNA model which represent DNA as beads in a chain with interaction potential derived by IMC from atomistic simulations including effects of ions and water, we simulated a very long (40,000 base pairs) DNA and found that it, in presence of $CoHex^{3+}$ ions, form toroidal structures which were also observed in electron microscopy studies (Sun et al. 2019) (Fig. 6).

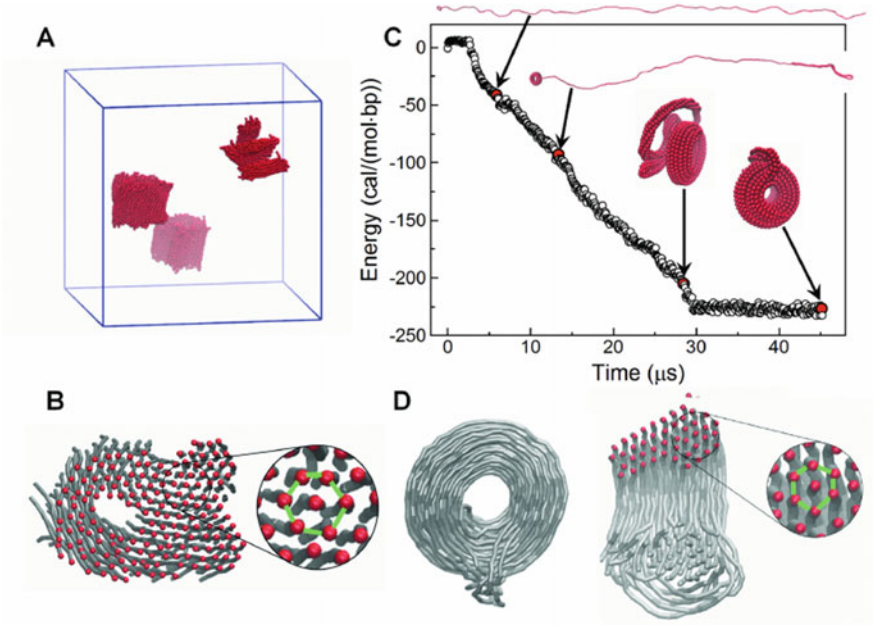

Fig. 6 DNA aggregation and toroid formation induced by CoHex3+ ions simulated by the super coarse-grained DNA model. **A**: Final configuration of DNA aggregation in a simulation with 400 DNA oligonucleotides of 100 base pairs each. **B**: Cross-section of one of the DNA condensed particle shown in **A**. **C** and **D**: Formation of toroidal structures in the super coarse-grained simulations of a 10,200 base pairs DNA. C: Energy profile and snapshots (normalised per DNA base pair) from one of the simulations. **D**: Structure of the DNA toroid. Cartoon on the right-hand side shows cross-section through the toroid where the red dots illustrate DNA double helices near the cutting plane. The zoom-in illustrated in **B** and **D** show the DNA packing within the aggregates with green lines highlighting hexagonal arrangement of the DNA molecules. *Reprinted* from Sun et al. (2019) with permission from Oxford University Press

4 Future Prospects

4.1 Chromosome Modelling

The hierarchical multiscale modelling of chromatin can be extended to reach what can be still considered as a grey-zone and to study the organization of the chromosomes inside the cell nucleus. This is a highly challenging "Bottom-up meets the Top-Down" project and is recognized as such by many colleagues (Langowski and Heermann 2007; Ozer et al. 2015; Moller and Pablo 2020; Bendandi et al. 2020; Caudai et al. 2019). In doing so modelling is not enough but we need to use a more integrative approach where we combine competences from other areas including bioinformatics (Ozer et al. 2015) and advanced instrumentation (Moller and Pablo 2020). Successively we also need to complement the purely physics and particle based methods by adding knowledge and rule-based models in the toolbox. We need experimental information about topologically associated domains (TAD) to insert in the models (Caudai et al. 2019).

We also will to study the role (structure and function) of nuclear lamina in the organisation of the chromosome, including the nuclear envelope proteins lamins (globular regulators of chromatin) bound to the lamina and other inner membrane proteins interacting with the chromosome (Dittmer and Misteli 2011; Ho and Lammerding 2012; Nora et al. 2013). Mutations in polymerizing lamins (laminopathies) are connected to a diversity of genetic human diseases (currently 17 are known) giving indication of their close role in cellular functions being important modulators in transcriptional regulation while affecting chromatin structure and organization. We will apply our multi-scale methodology to these higher-order polymeric proteins with alpha-helical rod-domains and study their roles as regulators of chromatin.

The function of the genes is not only what is encoded linearly in the sequences but it is now known that the topological organization of the DNA in chromatin and how it interacts with the nuclear environment are important in transcriptional regulation. Also, chromatin may have different structures inside the cell nucleus depending of the cell cycle. Chromosome Conformation Capture (3C) and its many extensions and other related methods have become highly valuable to decipher the spatial organisation of chromatin although they do not give information about the dynamical processes (Emmett et al. 2015; Lieberman-Aiden et al. 2009). Local DNA contacts are important in communications between the enhancers and promoters and these take place in TADs which in turn shape the chromosome landscape. 3C methods are used to quantify the number of interaction between the chromosome positions. Among the most popular of the 3C methods currently is the Hi-C (Reddy and Feinberg 2013) which can also combined with other methods, for example three-dimensional fluorescence in situ hybridization (3D-FISH) (https://www.ncbi.nlm.nih.gov/geo/). This method gives a picture and behaviour of packed chromosome as fractal globule. The Hi-C (extension of 3C methods) is becoming more accurate due to better regression-based correction schemes.

4.2 Genome-Wide Modeling

Hi-C maps (available for example from Gene Expression Omnibus (Ibrahim et al. 2013) can be converted into a matrix of average pairwise distances that can be used as restraints in many simulation models. This is analogous to when NOE distances are added to all-atom force fields in determination of protein structures using NMR. This allows for the first time a genome-wide modelling for example by using Brownian Dynamics simulations. Currently restraint-base modelling has been performed for chromatin using (1) rule-based spatial modelling (Doyle et al. 2014; Imakaev et al. 2015; Serra et al. 2015), (2) polymer models (Gruenert et al. 2010; Carstens et al. 2016) and (3) bead models (Szałaj et al. 2016). The resolution of the used models depends on the resolution in the genome data sets. However it is increasing continuously. As an example of a bead model with resolution of 50kbp requires 3330 beads to represent the X chromosome (Szałaj et al. 2016). The parameterization of the bead models is very much ad hoc with input from experiments, while we wish to be able to use super-coarse-graining based on underlying molecular information. Genome-wide modelling is becoming a very vital area. To become more acquainted we suggest following excellent reviews for multi-scale modelling and other new approaches (Molitor et al. 2017; Xia 2018; Sewitz et al. 2017; Bsascom and Schlick 2017).

5 Final Remarks

In this Chapter we introduce and discuss "Inverse Problem" in Chemistry and focus on its applications in Computational Chemistry where we can produce models for interacting molecular systems (interaction potentials and force fields) backward from the already obtained simulation results. We demonstrate that by proceeding systematically and hierarchically this inverse procedure can be made through hierarchical coarse-graining to a powerful and accurate multiscale modelling methodology called the Inverse Monte Carlo (IMC) alternatively Newtonian Inversion, which can be used all the way from first-principles quantum mechanics to meso-scale and beyond meaning in such a way that no empirical information would be needed. We do not do this yet routinely as computers need to become much faster for first-principles simulations of biological material as an example. We start normally from all-atom classical simulations. We discuss both the benefits and limitations of the IMC methods while illustrating some important applications, especially applications where the method is clearly superior to other similar methods in its accuracy. We show examples of higher-order coarse-graining (super-coarse-graining). We then discuss future aspects of the method and introduce the area of genome-wide modelling where we expect to make an impact with IMC combined informatics and knowledge/rule-based methods including machine learning and closely carried out with several sophisticated experimental techniques currently used in genome research.

Acknowledgements A. Laaksonen acknowledges Swedish Research Council for financial support (project: 2019-03865), and partial support from a grant from Ministry of Research and Innovation of Romania (CNCS-UEFISCDI, project number PN-III-P4-ID-PCCF-2016-0050, within PNCDI III). A. Lyubartsev acknowledges Swedish Research Council (project: 2017-03950) and Horizon 2020 framework programme (SmartNanoTox project) for financial support.

Conflict of Interest The authors declare no conflict of interest.

Ethical Aproval The authors declare that this article does not contain any studies with human participants or animals.

References

Bascom G, Schlick T (2017) Linking chromatin fibers to gene folding by hierarchical looping. Biophys J 112:434–445

Bendandi A, Dante S, Zia SR, Diaspro A, Rocchia W (2020) Chromatin compaction multiscale modeling: a complex synergy between theory, simulation, and experiment. Front Mol Biosci 7:1–14

Brandt EG, Lyubartsev AP (2015) Molecular dynamics simulations of adsorption of amino acid side chain analogues and a titanium binding peptide on the TiO_2 (100) surface. J Phys Chem C 119:18126–18139

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. Nature 559:547–555

Carstens S, Nilges M, Habeck M (2016) Inferential structure determination of chromosomes from single-cell Hi-C data. PLoS Comput Biol 12:e1005292

Caudai C, Salerno E, Zoppe M, Tonazzini A (2019) Estimation of the spatial chromatin structure based on a multiresolution bead-chain model. IEEE/ACM Trans Comput Biol Bioinform 16:550–559

Cherni A, Piersanti E, Anthoine S, Chaux C, Shintu L, Yemloul M, Torresani B (2019) Challenges in the decomposition of 2D NMR spectra of mixtures of small molecules. Faraday Discuss 218:459–480

Cho HM, Chu J-W (2009) Inversion of radial distribution functions to pair forces by solving the Yvon-Born-Green equation iteratively. J Chem Phys 131:134107

Dans PD, Walther J, Gomez H, Orozco M (2016) Multiscale simulation of DNA. Curr Opin Struct Biol 37:29–45

Dittmer T, Misteli T (2011) The lamin protein family. Genome Biol 12:222

Doyle B, Fudenberg G, Imakaev M, Mirny LA (2014) Chromatin loops as allosteric modulators of enhancer-promoter interactions. PLoS Comput Biol 10:e1003867

K Emmett, B Schweinhart, R Rabadan (2015) Multiscale topology of chromatin folding. In: Proceedings of the 9th EAI international conference on bio-inspired information and communications technologies (formerly BIONETICS), BICT 2015, pp 177–180

Ercolessi F, Adams JB (1994) Interatomic potentials from first-principles calculations: the force-matching method. Europhys Lett 26(8):583–588

Fan Y, Korolev N, Lyubartsev AP, Nordenskiöld L (2013) An advanced coarse-grained nucleosome core particle model for computer simulations of nucleosome-nucleosome interactions under varying ionic conditions. PLoS ONE 8:e54228

Ferguson AL (2018) Machine learning and data science in soft materials engineering. J Phys Condens Matter 30:1–27

Franceschetti A, Zunger A (1999) The inverse band-structure problem of finding an atomic configuration with given electronic properties. Nature 402:60–63

Gruenert G, Ibrahim B, Lenser T, Lohel M, Hinze T, Dittrich P (2010) Research article rule-based spatial modeling with diffusing, geometrically constrained molecules. BMC Bioinform 11:307

Hachmann J, Afzal MAF, Haghighatlari M, Pal Y (2018) Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space. Mol Simul 44:921–929

Halappanavar S, van den Brule S, Nymark P, Gaté L, Seidel C, Valentino S, Zhernovkov V, Danielsen P, De Vizcaya A, Wolff H, Stöger T, Boyadziev A, Poulsen SS, Sørli JB, Vogel U (2020) Adverse outcome pathways as a tool for the design of testing strategies to support the safety assessment of emerging advanced materials at the nanoscale. Part Fibre Toxicol 17. Article number: 16

Henderson RL (1974) A uniqueness theorem for fluid pair correlation functions. Phys Lett 49A:197–198

Ho CY, Lammerding J (2012) Lamins at a glance. J Cell Sci 125:2087–2093

https://www.ncbi.nlm.nih.gov/geo/

Hu XQ, Beratan DN, Yang WT (2009) Emergent strategies for inverse molecular design. Sci China Ser B Chem 52:1769–1776

Ibrahim B, Henze R, Gruenert G, Egbert M, Huwald J, Dittrich P (2013) Spatial rule-based modeling: a method and its application to the human mitotic kinetochore. Cells 3:506–544

Imakaev MV, Fudenberg G, Mirny LA (2015) Modeling chromosomes: beyond pretty pictures. FEBS Lett 589:3031–3036

Izvekov S, Voth GA (2005) Multiscale coarse graining of liquid-state systems. J Chem Phys 123:134105

Johnson MD, March NH (1963) Long-range oscillatory interaction between ions in liquid metals. Phys Lett 4:313–314

Jonas E (2019) Deep imitation learning for molecular inverse problems. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada

Jones JE (1924) On the determination of molecular fields. II. From the equation of state of a gas. Proc Roy Soc Lond Ser A 106:463–477

Korolev N, Lyubartsev AP, Laaksonen A (2004) Electrostatic background of chromatin fiber stretching. J Biomol Struct Dyn 22:215–226

Korolev N, Lyubartsev AP, Nordenskiöld L (2006) Computer modeling demonstrates that electrostatic attraction of nucleosomal DNA is mediated by histone tails. Biophys J 90:4305–4316

Korolev N, Allahverdi A, Yang Y, Fan Y, Lyubartsev AP, Nordenskiöld L (2010) Electrostatic origin of salt-induced nucleosome array compaction. Biophys J 99:1896–1905

Kunkin W, Frisch HL (1969) Inverse problem in classical statistical mechanics. Phys Rev 177:282–287

Langowski J, Heermann DW (2007) Computational modeling of the chromatin fiber. Semin Cell Dev Biol 18:659–667

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–294

Lopez H, Brandt EG, Mirzoev A, Zhurkin D, Lyubartsev A, Lobaskin V (2017) Multiscale modelling of bionano interface. In: Tran L et al (eds) Modelling the toxicity of nanoparticles. Advances in experimental medicine and biology, vol 947

Lyubartsev AP, Laaksonen A (1995) Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. Phys Rev E 52:3730–3737

Lyubartsev AP, Laaksonen A (1996) Concentration effects in aqueous NaCl solutions. A molecular dynamics simulation. J Phys Chem 100:16410–16418

Lyubartsev AP, Laaksonen A (1997) Osmotic and activity coefficients from effective potentials for hydrated ions. Phys Rev E 55:5689–5696

Lyubartsev A, Laaksonen A (1998) Molecular dynamics simulations of DNA in solution with different counter-ions. J Biomol Struct Dyn 16(3):579–592

Lyubartsev AP, Laaksonen A (2000) Determination of effective pair potentials from *ab initio* simulations: application to liquid water. Chem Phys Lett 325(1–3):15–21
Lyubartsev AP, Laaksonen A (1999) Effective potentials for ion-DNA interactions. J Chem Phys 111(24):11207–11215
Lyubartsev AP, Laasonen K, Laaksonen A (2001) Hydration of Li+ ion. An ab initio molecular dynamics simulation. J Chem Phys 114(7):3120–3126
Lyubartsev A, Tu Y, Laaksonen A (2009) Hierarchical multiscale modelling scheme from first principles to mesoscale. J Theor Comput Nanosci 6(5):951–959
Lyubartsev A, Mirzoev A, Chen LJ, Laaksonen A (2010) Systematic coarse-graining of molecular models by the Newton inversion method. Faraday Discuss 144:43–56
Lyubartsev A, Naome A, Laaksonen A (2015) Systematic Hierarchical Coarse-Graining with the Inverse Monte Carlo Method. J Chem Phys 143(24):243120/1-243120/8
Martinez-Luaces V (2012) Chemical kinetics and inverse modelling problems. In: Patel V (ed) Chemical Kinetics. Intech Publishing
Mechelke M, Habeck M (2013) Estimation of interaction potentials through the configurational temperature formalism. J Chem Theory Comput 9:5685–5692
Mirzoev A, Lyubartsev AP (2011) Effective solvent mediated potentials of Na+ and Cl− ions in aqueous solution: temperature dependence. Phys Chem Chem Phys 13:5722–5727
Mirzoev A, Lyubartsev AP (2013) MagiC: software package for multiscale modeling. J Chem Theory Comput 9:1512–1520
Mirzoev A, Nordenskiöld L, Lyubartsev A (2019) Magic v.3: an integrated software package for systematic structure-based coarse-graining. Comput Phys Commun 237:263–273. Software MagiC can be downloaded from: https://www.fos.su.se/~sasha/magic/
Molitor J, Mallm J-P, Rippe K, Erdel F (2017) retrieving chromatin patterns from deep sequencing data using correlation functions. Biophys J 112:473–490
Moller J, de Pablo JJ (2020) Bottom-up meets top-down: the crossroads of multiscale chromatin modeling. Biophys J 118:2057–2065
Mullinax JW, Noid WG (2010) A generalized-Yvon-Born-Green theory for determining coarse-grained interaction potentials. J Phys Chem C 114:5661–5674
Naômé A, Laaksonen A, Vercauteren DP (2014) A solvent-mediated coarse-grained model of DNA derived with the systematic Newton inversion method. J Chem Theory Comput 10(8):3541–4354
Naômé A, Laaksonen A, Vercauteren D (2015) A CG simulation study of the structures, energetics, and dynamics of linear and circular DNA with its ions. J Chem Theory Comput 11:2813–2826
Noh J, Gu GH, Kim S, Jung Y (2020) Machine-enabled inverse design of inorganic solid materials: promises and challenges. Chem Sci 11:4871–4881
Nora EP, Dekker J, Heard E (2013) Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? BioEssays 35:818–828
Ozer G, Luque A, Schlick T (2015) The chromatin fiber: multiscale problems and approaches. Curr Opin Struct Biol 31:124–139
Plimpton S (1995) Fast parallel algorithms for short-range molecular dynamics. J Comput Phys 117:1–19
Pombo A, Dillon N (2015) Three-dimensional genome architecture: players and mechanisms. Nat Rev 16:245–257
Power D, Rouse I, Poggio S, Brandt E, Lopez H, Lyubartsev A, Lobaskin V (2019) A multiscale model of protein adsorption on a nanoparticle surface. Model Simul Mater Sci Eng 27:084003
Reatto L, Levesque D, Weis JJ (1986) Iterative predictor-corrector method for extraction of the pair interaction from structural data for dense classical liquids. Phys Rev A 33:3451–3465
Rebic M, Mocci F, Laaksonen A, Ulicny J (2015) Multiscale simulations of human telomeric G-quadruplex DNA. J Phys Chem B 119(1):105–113
Rebic M, Mocci F, Ulicny J, Lyubartsev A, Laaksonen A (2017) Coarse-grained simulation of rodlike higher-order quadruplex structures at different salt concentrations. ACS Omega 2(2):386–396
Reddy KL, Feinberg AP (2013) Higher order chromatin organization in cancer. Semin Cancer Biol 23:109–115

Reith D, Putz M, Muller-Plathe F (2003) Deriving effective mesoscale potentials from atomistic simulations. J Comput Chem 24:1625–1636
Rosenfeld Y, Kahl G (1997) The inverse problem for simple classical liquids: a density functional approach. J Phys Condens Matter 9:L89–L98
Sanchez-Lengeling B, Aspuru-Guzik A (2018) Inverse molecular design using machine learning: generative models for matter engineering. Science 361:360–365
Schommers W (1983) Pair potentials in disordered many-particle systems: a study for liquid gallium. Phys Rev A 28:3599–3605
Schwarz-Plaschg C, Kallhoff A, Eisenberger I (2017) Making nanomaterials safer by design? Nanoethics 11:277–281
Serra F, Di Stefano M, Spill YG, Cuartero Y, Goodstadt M, Baù D, Marti-Renom MA (2015) Restraint-based three-dimensional modeling of genomes and genomic domains. FEBS Lett 589:2987–2995
Sewitz SA, Fahmi Z, Lipkow K (2017) Higher order assembly: folding the chromosome. Curr Opin Struct Biol 42:162–168
Shell MS (2008) The relative entropy is fundamental to multiscale and inverse thermodynamic problems. J Chem Phys 129:144108
Sherman ZM, Howard MP, Lindquist BA, Jadrich RB, Truskett TM (2020) Inverse methods for design of soft materials. J Chem Phys 152:140902
Soper AK (1996) Empirical potential Monte Carlo simulation of fluid structure. Chem Phys 202:295–306
Struebing H, Ganase Z, Karamertzanis PG, Siougkrou E, Haycock P, Piccione PM, Armstrong A, Galindo A, Adjiman CS (2013) Computer-aided molecular design of solvents for accelerated reaction kinetics. Nature Chem 5:952–957
Sun T, Mirzoev A, Minhas V, Korolev N, Lyubartsev AP, Nordenskiöld L (2019) A multiscale analysis of DNA phase separation: from atomistic to mesoscale level. Nucleic Acids Res 47:5550–5562
Swendsen RH (1979) Monte Carlo renormalization group. Phys Rev Lett 42:859–861
Szałaj P, Tang Z, Michalski P, Pietal MJ, Luo OJ, Sadowski M, Li X, Radew K, Ruan Y, Plewczynski D (2016) An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. Genome Res 26:1697–1709
Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. SIAM, Philadephia. ISBN 0-89871-572-5
Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718
van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES (2010) Hi-C: a method to study the three-dimensional architecture of genomes. JOVE J Vis Exp 39:e1869
von Lilienfeld OA, Lins RD, Rothlisberger U (2005) Variational particle number approach for rational compound design. Phys Rev Lett 95:153002
Wang M, Hu X, Beratan DN, Yang W (2006) Designing molecules by optimizing potentials. J Am Chem Soc 128:3228–3232
Wang Y-L, Lyubartsev A, Lu Z-Y, Laaksonen A (2013) Effect of the electrostatic interactions on the multiscale coarse-graining simulations of ionic liquids. Phys Chem Chem Phys 16:1–12
Xia K (2018) Sequence-based multiscale modeling for high-throughput chromosome conformation capture (Hi-C) data analysis. PLoS ONE 13:e0191899
Yaman F, Yakhno VG, Potthast R (2013) A survey on inverse problems for applied sciences. Math Probl Eng 1–19. Article ID: 976837
Zhu Y-L, Liu H, Li Z-W, Qian H-J, Milano G, Lu Z-Y (2013) GALAMOST: GPU-accelerated large-scale molecular simulation toolkit. J Comput Chem 34:2197–2211