Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Neurocomputing 56 (2004) 365–379

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Topology representing neural networks reconcile biomolecular shape, structure, and dynamics

Willy Wriggers[a,*], Pablo Chacón[a], Julio A. Kovacs[a], Florence Tama[a], Stefan Birmanns[b]

[a] *Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA*

[b] *Central Institute for Applied Mathematics, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany*

Received 3 August 2002; received in revised form 20 May 2003; accepted 22 September 2003

Abstract

Topology-representing networks (TRNs) generate reduced models of biomolecules and thereby facilitate the fitting of molecular fragments into large macromolecular complexes. The components of such complexes undergo a wide range of motions, and shapes observed at low resolution often deviate from the known atomic structures. What is required for the modeling of such motions is a combination of global shape constraints based on the low-resolution data with a local modeling of atomic interactions. We present a novel Motion Capture Network that freezes inessential degrees of freedom to maintain the stereochemistry of an atomic model. TRN-based deformable models retain much of the mechanical properties of biological macromolecules. The elastic models yield a decomposition of the predicted motion into vibrational normal modes and are amenable to interactive manipulation with haptic rendering software.

Keywords: Volumetric registration; Haptic rendering; Macromolecular assemblies; Multi-resolution docking; Normal mode analysis

1. Introduction

Scientific computing had a profound influence on the historic success of structural biology. More than 23,000 biomolecular structures are known today at atomic resolution

* Corresponding author. Current address: School of Health Information Sciences, University of Texas - Houston, 7000 Fannin, Houston, TX 77030, USA. Tel.: +1(713)500-3961; fax: +1(713)500-3907.
E-mail address: wriggers@biomachina.org (W. Wriggers).

0925-2312/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2003.09.007

owing to the advancement of numerical algorithms and computational speed that facilitated the data processing of NMR and X-ray crystallographic data. Likewise, electron microscopy (EM) benefited tremendously from image processing methods that were developed in the engineering sciences since the 1950s. The underlying physical concepts of concurrent structural biology software, diffraction theory and spectroscopy, are well understood and the corresponding algorithmic developments in structural biology have by now reached a high level of maturity. The availability of databases and vast amounts of structure information recently have prompted new trends in structure-based computing that are more concerned with the mining, as opposed to creating, of the biophysical data.

Structural bioinformatics involves the statistical analysis and architecture of biomolecular structures at multiple levels of resolution, from the atomic scale to low-resolution EM image reconstructions. Modern information processing techniques, such as artificial neural networks, in concert with physics-based classical simulation methods, combine structural data from a variety of biophysical sources: X-ray crystallography, EM, small-angle X-ray scattering, fluorescence spectroscopy, and biochemical labeling and footprinting. In this paper, we describe the use of topology-representing networks (TRNs) for combining structural data from a variety of biophysical origins.

Current advances in biology and medicine depend on an understanding of fundamental cellular processes, most of which involve the actions and interactions of large biomolecular assemblies of mega-Dalton molecular weight. Three-dimensional (3D) structures and image reconstructions of assemblies, involving hundreds of thousands to millions of atoms, are now routinely determined by X-ray crystallography and cryo-EM [14,36]. Nearly every major process in a cell is carried out by assemblies of 10 or more biomolecules [2]. Cytoskeletal filaments such as actin, symmetric assemblies such as chaperonins and viruses, as well as the ribosome, spliceosome, and RNA polymerase complexes, are highly evolved macromolecular assemblies comprised of many protein and nucleic acid subunits.

Medium resolution modeling constitutes a promising path to the simulation of large biomolecular assemblies. In the past 3 years we have developed a novel technology that enables a tessellation of both atomic resolution structures and low-resolution data from EM. In Section 2, we review the TRN algorithm as it is applied to multi-resolution biophysical data. Our unsupervised learning approach differs from the work of other authors in the bioinformatics field who have used *supervised* techniques for the characterization of biological data with artificial neural networks [39,33].

In Section 3, we use TRNs for the rigid-body fitting with a force-feedback device and derive the equations for accurate force and torque calculations that assist an expert user in the model building in a virtual reality environment. The developed algorithms are utilized in molecular visualization routines. This effort will permit scientists to build models interactively within a single computational environment.

In Section 4, we describe for the first time the algorithmic details of a novel flexible fitting algorithm, the Motion Capture Network (MCN). The term "Motion Capture" suggests an analogy to the technology of the same name in the entertainment industry and in biomechanics, where human-like motion is captured and digitized by fitting trussed networks (skeletons) to the positions of human extremities recorded from

visually tracked actors. In biomolecular applications the skeleton-based fitting approach provides robustness against the effects of noise and experimental uncertainty that would otherwise lead to significant local distortions in the flexed protein models.

Normal mode analysis [9,10] (NMA) involves the decomposition of the flexing motion into vibrational modes based on an elastic model. In Section 5, we will parametrize our TRN-based deformable models such that the NMA-derived motions optimally approximate the motions of atomic structures. The goal is to maintain continuity of elastic models at all resolution scales.

2. TRNs capture 3D structures at a reduced level of detail

The algorithms described in this section allow one to discretize both high- and low-resolution biological data by a small number of *neural pointers* (also known in the literature as *fiducials*, *codebook vectors*, *feature points*, or *landmarks*) that characterize the shape and density distribution of the occupied volume. A large variety of clustering techniques exist that represent data at reduced spatial resolution. Vector quantization [16], in particular, has been developed since the 1950s as a tool for speech and image compression. One of the requirements for our work in 3D registration is the statistical reproducibility of the found neural pointers, which limits the number of suitable methods. We currently favor TRNs due to their desirable convergence properties [28,46], as described in the following.

Let us assume a set of neurons with weights $\mathbf{w}_i$ ($i=1,\ldots,K$). Furthermore, we assume that each neuron receives the same external input signals $\mathbf{v} \in M \subset \Re^D$. The signals $\mathbf{v}$ will be randomly selected on the manifold M according to a probability density function $P(\mathbf{v})$ (e.g. corresponding to the atomic masses or low-resolution density). The adaptation of the $\mathbf{w}_i$ to the input signals is affected by the topological arrangement of the *Voronoi cells* V_i, defined by

$$V_i = \{\mathbf{u} \in \Re^D \mid \|\mathbf{u}-\mathbf{w}_i\| \leqslant \|\mathbf{u}-\mathbf{w}_j\|,\ j=1,\ldots,K\},\ i=1,\ldots,K. \tag{1}$$

Fig. 1a shows Voronoi cells in $\Re^2$ corresponding to 18 neurons in input space. In $\Re^3$, the cell boundaries are subsets of the bisecting planes between each pair of neurons adjacent in input space.

Information about the arrangement of the Voronoi cells is given by the closeness rank λ_i of each neural pointer depending on $\mathbf{v}$, i.e. the number of neurons $\mathbf{w}_j$ with $\|\mathbf{v}-\mathbf{w}_j\| < \|\mathbf{v}-\mathbf{w}_i\|$. The adaptation of the $\mathbf{w}_i$ at a given time step t, $t=1,\ldots,t_{\max}$, is given by

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \varepsilon \cdot \mathrm{e}^{-\lambda_i/\sigma}(\mathbf{v}-\mathbf{w}_i(t)), \tag{2}$$

where the neuron plasticity ε and the proximity width σ are monotonically decreasing with compute time according to $\sigma = \sigma_i(\sigma_f/\sigma_i)^{t/t_{\max}}$, and $\varepsilon = \varepsilon_i(\varepsilon_f/\varepsilon_i)^{t/t_{\max}}$.

One can show [24,16] that the limiting case ($\sigma \to 0$) corresponds to stochastic gradient descent minimization of the *encoding distortion error* that measures the mean-square deviation of the data from the neural pointers:

$$E = \int \|\mathbf{v}-\mathbf{w}_{i(\mathbf{v})}\|^2 P(\mathbf{v})\,\mathrm{d}\mathbf{v}, \tag{3}$$

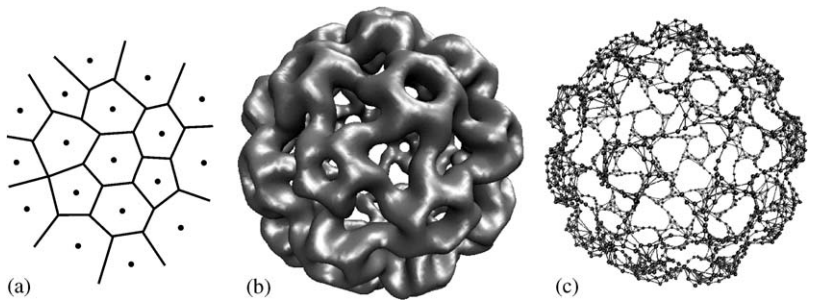

Fig. 1. Functionality of Voronoi tessellation, TRN, and competitive Hebb rule. (a) Tessellation of 2D space into 18 Voronoi cells. (b) The swollen form of cowpea chlorotic mottle virus at 23 Å resolution [35]. (c) Tessellation of the density with the TRN method (Eq. (2)) using 1380 neural pointers (shown as beads). The lateral connections were constructed using the competitive Hebb rule (see text). The following empirical parameters (Eq. (2)) were determined empirically: $\varepsilon_i = 0.1$; $\varepsilon_f = 0.001$; $\sigma_i = 276$; $\sigma_f = 0.02$, $t_{\max} = 1,000,000$. The 3D scenes in all figures were rendered with Situs [42] and the molecular graphics program VMD [23].

where $\lambda_{i(\mathbf{v})}(\mathbf{v}) = 0$. By means of the closeness ranking in Eq. (2) (i.e., $\sigma > 0$), TRN eludes the local minima of E during early training [28]. Ultimately, though, the annealing parameter σ vanishes (i.e., $e^{-\lambda_i/\sigma} \to \delta_{0,\lambda_i}$), and only the "winning" neural pointer ($\lambda_i = 0$) is updated at each step, which promotes that TRN ultimately settles at (or near) the global minimum of E. Fig. 1b and c presents a TRN tessellation of the 3D density of a virus capsid.

The main advantage of a TRN relative to the more widely known Kohonen [24] self-organizing map (SOM) is that the final distribution of the pointers is independent of a priori lateral neural connectivities. Rather, proximity relationships can be learnt from the distribution of neural pointers. The *competitive Hebb rule* [29] constructs connections between adjacent neurons, if the connection is, at least partially, covered by the density distribution $P(\mathbf{v})$. The algorithm is based on the mathematical theory of Delaunay triangulation [13]. At a given time step t, connections are formed between the two neural pointers closest to $\mathbf{v}$. The resulting connectivity structure defines a discrete topology- and path-preserving representation of M, even in cases where M has an intricate topology [28]. The competitive Hebb rule also defines "adjacency" between neural pointers in a mathematically consistent way. More theoretical details are given in Refs. [29,28]. An example of the application of this rule to 3D biophysical data is shown in Fig. 1c.

By repeating the TRN optimization (Eq. (2)) a number of times with statistically independent start positions that are randomly distributed according to $P(\mathbf{v})$ on the manifold M, one can estimate the convergence properties of the algorithm based on the resulting statistical variability of the neuron positions. Our studies revealed that the positions achieved with gradient descent ($\sigma \to 0$) were too unreliable to serve as markers for the docking of structures. In contrast, the TRN algorithm with the empirical parameters given in the caption of Fig. 1 is capable of very low variabilities on the order of an Angstrom, which is sufficiently precise for biomolecular docking. Furthermore, owing to its nature to describe the convergence towards a global optimum,

the averaged variability was useful for the estimation of the optimum complexity K of the network that results in the lowest spread in neural positions [42].

Reduced TRN-based models were first used in structural bioinformatics for rigid-body docking of atomic structures to low-resolution data, where an alignment of the data sets is achieved by identifying pairs of corresponding neurons from distance or connectivity matrices [44–46,15]. In the following, we describe a novel matching strategy based on TRNs and correlation functions that are suitable for interactive fitting.

3. Rigid-body docking and force-feedback

The quality of the match between a probe density $\rho_{\text{probe}}(\mathbf{r})$ and a target density $\rho_{\text{target}}(\mathbf{r})$ can be described by a correlation function:

$$C(\mathbf{R},\mathbf{T}) = \int \rho_{\text{probe}}(\mathbf{r},\mathbf{R},\mathbf{T})\rho_{\text{target}}(\mathbf{r})\,\mathrm{d}\mathbf{r}, \tag{4}$$

where $\mathbf{R}$ denotes the three rotational, and $\mathbf{T}$ the three translational degrees of freedom, respectively. In rigid-body fitting one would seek to maximize C, ideally by performing a full exploration of the 6D search space (Fig. 2a).

In addition to automated rigid-body fitting, microscopists have a need to evaluate and to manipulate docking models interactively "by eye". One of the challenges in

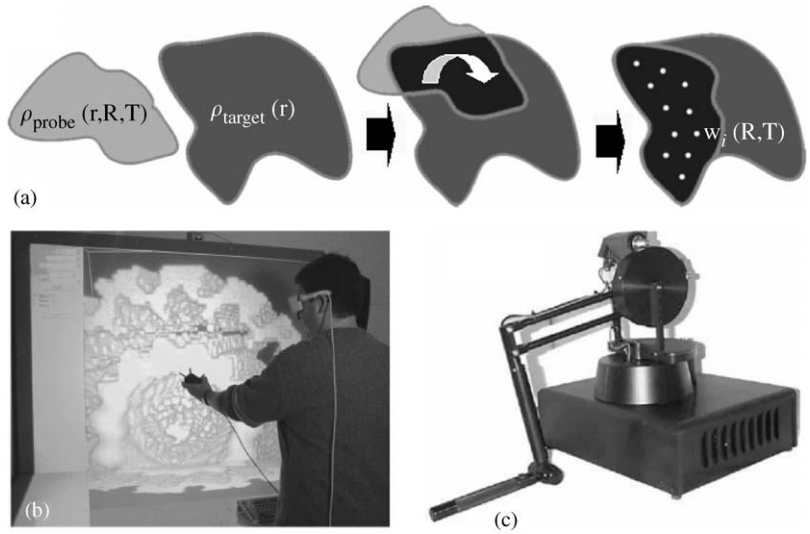

Fig. 2. Overview of rigid-body docking with TRNs and force-feedback. (a) Schematic diagram depicting the registration of a probe density $\rho_{\text{probe}}(\mathbf{r},\mathbf{R},\mathbf{T})$, subject to three translational ($\mathbf{T}$) and three rotational ($\mathbf{R}$) degrees of freedom, to a target density $\rho_{\text{target}}(\mathbf{r})$. The probe density can be approximated by a TRN with neurons $\mathbf{w}_i(\mathbf{R},\mathbf{T})$ for force and torque calculations (see text). (b) Visualization of a microtubule surface in an immersive VR environment using volslice3d [42], an earlier prototype of SenSitus. (c) The PHANTOM 1.5/6DOF force feedback device from SensAble Corp.

structural bioinformatics is to enable the efficient use and inter-operation of a diverse set of techniques to simulate, analyze, model, and visualize the complex architecture and interactions of macromolecular systems. To meet this challenge we develop a molecular graphics package termed "SenSitus" that is capable of supporting virtual reality (VR) devices such as stereo glasses, 3D trackers, and force-feedback (haptic) devices (Figs. 2b and c). Three-dimensional capabilities and the "physics of touch" offer tangible benefits for modelers who wish to explore a variety of docking situations in a VR environment (Fig. 2b). Our software supports this by calculating forces according to the correlation coefficient C. The high sampling frequency required for force feedback (refresh rate > 1 kHz) is achieved by means of the TRN algorithm that reduces the complexity of the data representation to manageable levels.

In the following we approximate the probe density by a sum of Dirac delta functions that are localized at the TRN neurons:

$$\rho_{\text{probe}}(\mathbf{r}, \mathbf{R}, \mathbf{T}) = \sum_{i=1}^{K} \delta(\mathbf{r} - \mathbf{w}_i(\mathbf{R}, \mathbf{T})). \tag{5}$$

The correlation function C is thereby reduced to a sum over the target density evaluated at the neuron positions:

$$C(\mathbf{R}, \mathbf{T}) = \sum_{i=1}^{K} \rho_{\text{target}}(\mathbf{w}_i(\mathbf{R}, \mathbf{T})). \tag{6}$$

Next, we define a potential energy $U = -\mu C$, where μ is a user-defined scaling factor. We seek to minimize U by interacting with the molecular data. The force $\mathbf{f}_i$ acting on an individual neuron i is the negative gradient of the potential energy, $\mathbf{f}_i = -\nabla_i U$. Therefore, the total force $\mathbf{F}$ acting on the centroid of the probe molecule is given by

$$\mathbf{F}(\mathbf{R}, \mathbf{T}) = \sum_{i=1}^{K} \mathbf{f}_i = \mu \sum_{i=1}^{K} \nabla \rho_{\text{target}}(\mathbf{w}_i(\mathbf{R}, \mathbf{T})). \tag{7}$$

Likewise, one can compute the total torque $\mathbf{Q}$ acting on the molecule. Shifting the origin of the neuron coordinate system to the centroid, we obtain

$$\mathbf{Q}(\mathbf{R}, \mathbf{T}) = \sum_{i=1}^{K} \mathbf{w}_i \times \mathbf{f}_i = \mu \sum_{i=1}^{K} \mathbf{w}_i \times \nabla \rho_{\text{target}}(\mathbf{w}_i(\mathbf{R}, \mathbf{T})). \tag{8}$$

The gradient field $\nabla \rho_{\text{target}}$ can be precomputed and is efficiently evaluated in real compute time by tri-linear interpolation.

The haptic device (Fig. 2c) measures a user's hand position and by means of Eqs. (7) and (8) exerts a precisely controlled force and torque on the hand. Therefore, the device not only enables the user to position and orient the probe structure relative to the target, but also directs the fitting to the next suitable location. This technique is very useful, as it facilitates the detection of possible fitting locations and simplifies the fine positioning of the structure.

One of the goals of our software development efforts is to allow researchers to build models, perform docking of atomic and volumetric data, visualize results from template convolution, and perform morphing and warping (flexible docking) interactively

within a single computational environment. Therefore, we need to devise a TRN that is designed for capturing conformations also in flexing situations where the probe and target molecule deviate from one another.

4. Motion capture networks: methodology and application to actin flexing

Rigid-body docking with TRNs, as implemented in our Situs docking package [45,42], laid the groundwork for the development of a flexible docking technique that brings deviating features of multi-resolution structures into register [41–43,12] if the atomic structure in one conformation is known. In such situations, the deviating atomic structure is moved towards the EM density by forcing the centroids of the Voronoi cells (Fig. 1a) of the atomic structure to coincide with the neural pointers of the EM density. This is done in a molecular dynamics refinement of the atomic structure where harmonic constraints between the Voronoi cell centroids and neural pointers form a global penalty that is imposed while preserving the moved structure at the local level [41]. The fitting accuracy that can be achieved in such flexible docking experiments is one order of magnitude above the nominal resolution of the EM map, or better [42].

One of the open questions in flexible docking, however, is how to maintain the stereochemical quality of a fitted structure, since any over-fitting to noisy experimental EM data would compromise the quality of the atomic model. Here, we describe for the first time the details of a significant improvement to our TRN-based flexible fitting algorithm, the Motion Capture Network (MCN). The basic idea is that lateral connections are formed between neurons that reflect the connectivity of the biological polypeptide chain. The resulting skeletons (distance-constrained lateral connections) eliminate the longitudinal degrees of freedom that are deemed inessential for the flexible docking, while permitting lateral flexibility. This approximation of the biomolecular motion can be justified by the statistics of biomolecular domain motions documented in the Protein Data Bank [1]. A significant majority (70%) of such observed motions can be classified either as hinge-bending or shearing motions. In both of these classes of motions, the longitudinal contributions to the conformational change (i.e. stretching or compression) are negligible compared to the lateral motions [18,17]. Only 7% of observed domain motions involve a partial refolding on the local level that would be difficult to predict by our reduced fitting to low-resolution data.

In the following we assume that we have tabulated a number of distance constraints, and the nth constraint between neurons $i(n)$ and $j(n)$ is given by

$$\mathbf{w}_{ij}^2 - d_{ij}^2 = 0, \tag{9}$$

where $\mathbf{w}_{ij} = \mathbf{w}_i - \mathbf{w}_j$, and d_{ij} is the desired spatial separation. These distance constraints are satisfied by adding displacements $\delta\mathbf{w}_i(t+1)$ to the neurons $\mathbf{w}_i(t+1)$ that resulted from an unconstrained TRN updating step (Eq. (2)). Formally, this problem is amenable to the Lagrangian formalism for holonomic constraints [20]. However, we consider here

a significantly simpler and more efficient approach for solving this problem iteratively:

$$\begin{aligned}\delta^n \mathbf{w}_i(t+1) &= g_{ij}\mathbf{w}_{ij}(t),\\ \delta^n \mathbf{w}_j(t+1) &= -g_{ij}\mathbf{w}_{ij}(t),\end{aligned} \tag{10}$$

where $\delta^n \mathbf{w}_i$ describes the action of the nth constraint and all constraints are satisfied in succession. This general ansatz, which can be motivated by the theory of numerical integration [34], treats each coefficient $g_{ij} = g_{ji}$ as an unknown that can be calculated. Consequently the unconstrained positions $\mathbf{w}_i(t+1)$ are corrected with $\sum_n \delta^n \mathbf{w}_i(t+1)$.

We define the partially corrected difference vector

$$\mathbf{w}'_{ij}(t+1) = \mathbf{w}_i(t+1) + \sum_{m<n} \delta^m \mathbf{w}_i(t+1) - \mathbf{w}_j(t+1) - \sum_{m<n} \delta^m \mathbf{w}_j(t+1), \tag{11}$$

the corresponding updating step

$$\delta\mathbf{w}_{ij}(t+1) = \delta^n \mathbf{w}_i(t+1) - \delta^n \mathbf{w}_j(t+1) \overset{(10)}{=} 2g_{ij}\mathbf{w}_{ij}(t) \tag{12}$$

and reformulate the constraints equation

$$(\mathbf{w}'_{ij}(t+1) + \delta\mathbf{w}_{ij}(t+1))^2 - d_{ij}^2 = 0, \tag{13}$$

which yields the quadratic equation for the unknown g_{ij}

$$4g_{ij}(\mathbf{w}_{ij}(t)\cdot\mathbf{w}'_{ij}(t+1)) + 4g_{ij}^2\mathbf{w}_{ij}^2(t) = d_{ij}^2 - \mathbf{w}'^2_{ij}(t+1). \tag{14}$$

This equation can be solved most efficiently to first order, i.e. the term in g_{ij}^2 may be neglected [31]. The enforcement of the nth constraint destroys to some degree all previous constraints $m < n$. Therefore, the algorithm is iterated in cyclic succession until the relative distortion $(\mathbf{w}_{ij}^2 - d_{ij}^2)/d_{ij}^2$ drops below a certain tolerance. For the MCN, we determined that a relative tolerance of 10^{-3} was sufficiently accurate and yielded a convergence within only 10–50 cycles through the constraints table. The iterative scheme should be initialized by setting $\mathbf{w}_i(1) = \mathbf{w}_i(0)$ $(i = 1,\ldots,K)$ before proceeding with the first TRN update (Eq. (2)).

The MCN was already tested in the flexing of RNA polymerase [43,12]. In particular, the formation of connectivities between neurons were depicted in Fig. 2 in [43]. The fitting accuracy that can be achieved with MCN is about one order of magnitude above the nominal resolution of the low-resolution density [42].

Fig. 3 shows an application of the MCN to the flexing of actin from a folded to the presumed unfolded (open) state of the protein. It had been shown biochemically [27] that actin's structural subdomains 3 and 4 (Fig. 3a) remain intact during the binding to the CCT chaperonin and the associated unfolding. We have thus chosen a level of detail ($K = 8$) in our reduced representation that would fully constrain subdomains 3 and 4 while affording some relative flexibility to subdomains 1 and 2. After assigning distance constraints among adjacent neurons that follow the polypeptide chain connectivity, the MCN was fitted to the open structure (Fig. 3b). Subsequently, the closed structure of actin was moved towards the open EM density by forcing the Voronoi cell centroids to coincide with the MCN neurons (Fig. 3c). This was done in a molecular dynamics refinement of the atomic structure with the Situs docking package

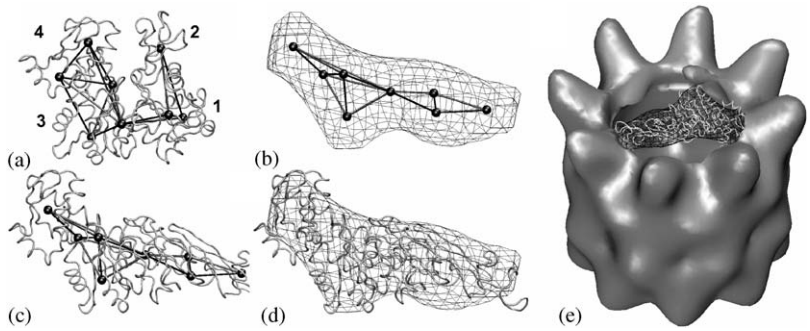

Fig. 3. Use of a MCN for the flexible docking of actin. (a) The atomic structure of actin's four subdomains in the closed conformation [47] is shown as a gray backbone trace. Eight neural pointers (shown as spheres) represent the actin structure. The distance-constrained connectivities are shown as black rods. (b) The density of open actin (shown as a wire mesh isocontour), extracted from the complex with CCT chaperonin [27]. The MCN of (a) was fitted to the density (see text). (c) The structure of actin after flexing based on the MCN displacements. (d) Comparison between flexed structure and density in the open form. (e) The densities [27] of the chaperonin CCT (gray solid isocontour) and of CCT-bound actin (black wire mesh isocontour) are shown with the flexibly fitted structure of actin (gray backbone trace).

[42] and X-PLOR [8]. The resulting model (Figs. 3d and e) provides a plausible hypothesis for the conformational change that can be tested experimentally. Due to the use of the reduced representation the structure does not appear overfitted even though all atomic degrees of freedom were considered in the flexing.

5. TRN-based deformable models

It is possible to obtain useful information on the dynamics, long-range coupling, and elastic properties of polynucleotides without requiring atomic resolution [48,40,7]. These studies suggest the importance of developing reduced structural models (Fig. 1) for large biomolecular assemblies to go beyond the size that can be handled at atomic detail. Normal mode analysis [9,10] (NMA), i.e. the decomposition of the motion into vibrational modes based on an elastic model of the biopolymer, is a frequently used technique to study the motion of large assemblies. Atomic motions corresponding to low-frequency normal modes are not localized [26]. Calculating the directional correlation functions, it was shown that NMA motions are highly correlated for atoms whose inter-atomic distances are within 5–10 Å [26]. Hence, it seems reasonable to expect that a sparse estimation of the displacement field using a TRN with a spatial resolution of 5–10 Å will reproduce the displacements well. After generating such a sparse estimation, displacements can be extended to the full space by interpolation. Moreover, a reduced description of the dynamics can be applied to both atomic structures and 3D image reconstructions from cryo-EM, as it is independent of the resolution of the underlying data.

The basic assumption (and limitation) of NMA is that the potential energy of the system varies quadratically about a given minimum energy conformation. This idea is rooted in the observation that biomolecules behave, more than expected, as if the energy surface were harmonic, even though the potential contains many local minima [22]. The methodology of NMA has already been discussed in many excellent textbooks and reviews [9,10,32], and we focus here on recent efforts to extend the method to the large systems of interest.

A first step in the reduction of the computational cost of NMA is the replacement of the atomic force field by a more simplified harmonic interaction potential of neighboring atoms. This approach, pioneered by Tirion [38], showed that low-frequency modes depend more on the global character of the deformations than on the precise form of the interaction potential. Still, at atomic resolution the standard Cartesian method involves a diagonalization of a $3N \times 3N$ matrix, where N is the number of atoms. The memory requirements are prohibitive for large proteins or assemblies with more than 500 residues. It is possible to reduce the degrees of freedom under consideration to the bond torsions [6,25] or by use of a Fourier basis [21]. However, for large assemblies it is more reasonable to reduce the amount of spatial detail in the model [11] while using the simplified harmonic interaction force field developed by Tirion:

We use the neural pointers that can be computed with the TRN algorithm (Fig. 1) for a reduced ($K \ll N$) representation of atomic structures or low-resolution data from cryo-EM. The pairwise Hookean potential between adjacent neural pointers is

$$U_{ij} = \frac{c}{2}(\|\mathbf{w}_{ij}\| - \|\mathbf{w}_{ij}^0\|)^2 = \frac{c}{2}\left(\frac{\mathbf{w}_{ij}^0 \cdot \Delta\mathbf{w}_{ij}}{\|\mathbf{w}_{ij}^0\|}\right)^2 + \mathrm{O}((\mathbf{w}_{ij}^0)^2), \tag{15}$$

where $\Delta\mathbf{w}_{ij} \equiv \mathbf{w}_{ij} - \mathbf{w}_{ij}^0$, and the zero superscript indicates the initial configuration. The strength of the potential c is an empirical constant for the system that can be adjusted such that the normal mode amplitudes match those from atomic detail NMA or to ensure consistency with experimentally observed properties.

The potential energy within the system is then given by

$$U = \sum_{i<j} U_{ij} C_{ij}, \tag{16}$$

where the connections C_{ij} are assigned within a certain distance cutoff [38,3] or learned with the competitive Hebb rule. It is straightforward to compute the $3K \times 3K$ Hessian matrix of second derivatives by expanding the U_{ij} to second order about $\mathbf{w}_{ij}^0$ (Eq. (15)). Since each neuron is assumed to have unit mass, the normal modes are the eigenvectors of the Hessian [10].

In [37] we have shown how NMA on a system of neural pointers connected with Hookean springs (Eq. (15)) can reproduce the experimentally observed atomic-resolution opening of the cleft in adenylate kinase in the reduced model at various levels of detail. Here, we demonstrate how essential motions can be extracted similarly from low-resolution cryo-EM maps. Fig. 4 compares the lowest-frequency mode from NMA with the motions from the flexible fitting of *T. aquaticus* RNA polymerase (RNAP) to a low-resolution map of *E. coli* RNAP [12]. The differences between the crystal and the cryo-EM isoforms (Fig. 4b) can be attributed to crystal packing effects

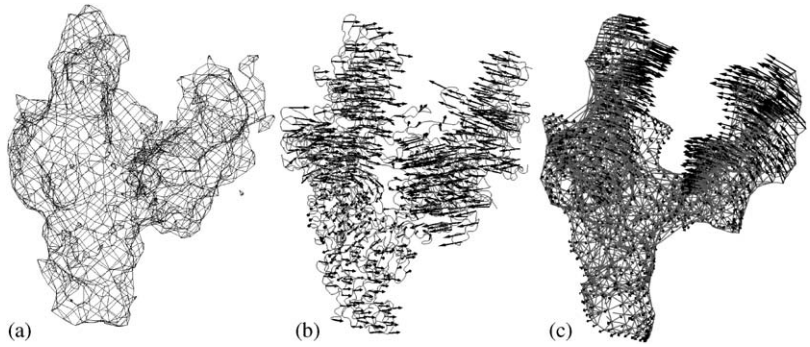

Fig. 4. Comparison of functionally relevant motions of RNA polymerase (RNAP) with TRN-based NMA. (a) Low-resolution cryo-EM map of *E. coli* RNAP as described [12]. (b) Displacements from flexible fitting of the *T. aquaticus* crystal structure to the cryo-EM density [12]. The arrows point from the flexibly fitted structure (backbone trace) to the original crystal conformation (not shown). (c) NMA (lowest-frequency mode) of the cryo-EM map using a TRN with $K = 1500$ neural pointers.

and reveal a closing of the RNAP jaws relative to the cryo-EM data. Interestingly, the lowest-frequency mode shows a very similar closing and opening motion of the jaws (Fig. 4c). The figure exemplifies how one could use NMA to predict functionally relevant motions from a single low-resolution structure.

6. Conclusions

In this paper, we have described a number of innovative neural network designs that facilitate the modeling and fitting of multi-resolution biophysical data in structural bioinformatics.

The TRN-based deformable models can be parametrized such that the NMA-derived motions optimally approximate the motions of small atomic structures sampled from molecular dynamics trajectories. The motions and conformational variability of atomic structures are sufficiently sampled by molecular dynamics simulations [31] provided that the systems are small. Quasi-harmonic analysis [4] is a statistical method related to NMA that determines the slow collective motions from a single simulation trajectory. However, for very large systems it is not possible to extract sufficient statistical information from molecular dynamics trajectories due to undersampling of large-scale displacements that would, e.g. lead to a severe overestimation of the stiffness of hinges between large domains. To this end we plan to model the elasticity of filaments, cross-bridges and coiled-coils with a vibrational analysis of TRNs. The goal of this work is to empirically fit the parameters in our models to reproduce experimentally measured mechanical properties such as flexural and torsional rigidity [5], stiffness [30] and the persistence length [19,5] of large protein assemblies.

The novel MCN implementation is available as part of our "classic" Situs distribution at http://situs.biomachina.org. At present the network complexity K and the distance constraints d_{ij} are user-defined parameters. It is desirable to automate the estimation

of these parameters for non-expert users of our software in the biological sciences. For example, the number K of neurons can be estimated by the number of independent pieces of information contained in a low-resolution EM reconstruction. This number can be obtained by dividing the total volume of the molecule by the volume of a single resolution element (i.e. a cube with a width corresponding to the nominal spatial resolution of the data). Also, one could automatically map the connectivity of the polypetide chain onto the reduced neuron representation by fitting a 1D Kohonen SOM [24], which represents the biomolecular backbone, to the TRN neurons that are interpreted as input space for the SOM. The implementation of such automated procedures are straightforward and subject of future revisions of our software.

SenSitus and the interactive force feedback fitting are already fully functional as a docking tool. The visualization program for various UNIX and PC architectures can be downloaded at http://sensitus.biomachina.org. Ultimately, we will integrate all of our advanced flexing functionality and the efficient simulation of multi-resolution data by normal modes analysis. The interactive flexing technology is currently in its infancy but it will undoubtedly gain in importance and popularity in the near future when more structures become available that require an induced fit of their components.

In summary, the new methods are adequate for the study of deformations and of dynamical properties of low-resolution biomolecular structures, or of very large structures, in which case simulations at the atomic level become prohibitively expensive.

Acknowledgements

We thank J.M. Valpuesta for kindly providing the EM maps of the apo and actin-bound CCT structures, and S. Darst for providing the *E. coli* RNA polymerase reconstruction. This work was supported by NIH grants P41-RR-12255 and 1R01-GM62968 and by the La Jolla Interfaces in Science Program/Burroughs Wellcome Fund.

References

[1] E.E. Abola, J.L. Sussman, J. Prilusky, N.O. Manning, Protein data bank archives of three-dimensional macromolecular structures, in: C.W. Carter Jr., R.M. Sweet (Eds.), Methods in Enzymology, Vol. 277, Academic Press, San Diego, 1997, pp. 556–571.

[2] B. Alberts, The cell as a collection of protein machines: preparing the next generation of molecular biologists, Cell 92 (1998) 291–294.

[3] I. Bahar, A.R. Atilgan, B. Erman, Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, Fold. Des. 2 (1997) 173–181.

[4] M.A. Balsera, W. Wriggers, Y. Oono, K. Schulten, Principal component analysis and long time protein dynamics, J. Phys. Chem. 100 (7) (1996) 2567–2572.

[5] D. ben Avraham, M.M. Tirion, Dynamic and elastic properties of F-actin: a normal-modes analysis, Biophys. J. 68 (1995) 1231–1245.

[6] B. Brooks, B. Karplus, Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor, Biophysics 80 (1983) 6571–6575.

[7] N. Bruant, D. Flatters, R. Lavery, D. Genest, From atomic to mesoscopic description of the internal dynamics of DNA, Biophys. J. 77 (1999) 2366–2376.

[8] A.T. Brünger, X-PLOR, Version 3.1: A System for X-ray Crystallography and NMR, The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, 1992.
[9] D.A. Case, Normal mode analysis of protein dynamics, Curr. Opin. Struct. Biol. 4 (1994) 285–290.
[10] D.A. Case, Normal mode analysis of biomolecular dynamics, in: W.F. van Gunsteren, P.K.Weiner, A.J. Wilkinson (Eds.), Computer Simulation of Biomolecular Systems, Vol. 3, Kluwer Academic Publishers, Dordrecht, Netherlands, 1997, pp. 284–301.
[11] P. Chacón, F. Tama, W. Wriggers, Mega-dalton biomolecular motion captured from electron microscopy reconstructions, J. Mol. Biol. 326 (2003) 485–492.
[12] S.A. Darst, N. Opalka, P. Chacón, A. Polyakov, C. Richter, G. Zhang, W. Wriggers, Conformational flexibility of bacterial RNA polymerase, Proc. Natl. Acad. Sci. USA 99 (2002) 4296–4301.
[13] B. Delaunay, Sur la spère vide, Bull. Acad. Sci. USSR VII (1934) 793–800.
[14] D.J. DeRosier, S.C. Harrison, Macromolecular assemblages: sizing things up, Curr. Opin. Struct. Biol. 7 (1997) 237–238.
[15] V.E. Galkin, A. Orlova, N. Lukoyanova, W. Wriggers, E.H. Egelman, ADF stabilizes an existing state of F-actin and can change the tilt of F-actin subunits, J. Cell Biol. 153 (2001) 75–86.
[16] A. Gersho, R.M. Gray, Vector quantization and signal compression, Kluwer Academic Publishers, Norwell, MA, 1992.
[17] M. Gerstein, W. Krebs, A database of macromolecular motions, Nucl. Acids Res. 26 (1998) 4280–4290.
[18] M. Gerstein, A.M. Lesk, C. Chothia, Structural mechanisms for domain movements in proteins, Biochemistry 33 (1994) 6739–6749.
[19] F. Gittes, B. Mickey, J. Nettleton, J. Howard, Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape, J. Cell Biol. 120 (1993) 923–934.
[20] H. Goldstein, Classical Mechanics, Addison-Wesley, Reading, MA, 1980.
[21] K. Hinsen, Analysis of domain motions by approximate normal mode calculations, Proteins: Struct. Funct. Genet. 33 (1998) 417–429.
[22] T. Horiuchi, N. Go, Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme, Proteins: Struct. Funct. Genet. 10 (1991) 106–116.
[23] W.F. Humphrey, A. Dalke, K. Schulten, VMD—visual molecular dynamics, J. Mol. Graphics 14 (1996) 33–38.
[24] T. Kohonen, Self-Organizing Maps, 2nd Edition, Springer, Berlin, 1995.
[25] M. Levitt, C. Sander, P.S. Stern, Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme, J. Mol. Biol. 181 (1985) 423–447.
[26] D. Lin, A. Matsumoto, N. Go, Normal mode analysis of a double-stranded DNA dodecamer d(CGCGAATTCGCG), J. Chem. Phys. 107 (1997) 3684–3690.
[27] O. Llorca, J. Martin-Benito, M. Ritco-Vosonvici, J. Grantham, G.M. Hynes, K.R. Willison, J.L. Carrascosa, J.M. Valpuesta, Eukaryotic chaperonin CCT stabilizes actin and tubulin folding intermediates in open quasi-native conformations, EMBO J. 19 (2000) 5971–5979.
[28] T.M. Martinetz, S.G. Berkovich, K. Schulten, "Neural gas" for vector quantization and its application to time-series prediction, IEEE Trans. Neural Networks 4 (4) (1993) 558–569.
[29] T. Martinetz, K. Schulten, Topology representing networks, Neural Networks 7 (1994) 507–522.
[30] A. Matsumoto, N. Go, Dynamic properties of double-stranded DNA by normal mode analysis, J. Chem. Phys. 110 (1999) 11070–11075.
[31] J.A. McCammon, S.C. Harvey, Dynamics of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1987.
[32] D.A. McQuarrie, Statistical Mechanics, Harper and Row, New York, 1976.
[33] K. Nakata, Prediction of zinc finger DNA binding protein, Comput. Appl. Biosci. 11 (1995) 125–131.
[34] J.-P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes, J. Comput. Phys. 23 (1977) 327–341.
[35] J.A. Speir, S. Munshi, G. Wang, T.S. Baker, J.E. Johnson, Structures of the native and the swollen forms of cowpea chlorotic mottle virus determined by X-ray crystallography and cryo-electron microscopy, Structure 3 (1995) 63–78.

[36] M.H.B. Stowell, A. Miyazawa, N. Unwin, Macromolecular structure determination by electron microscopy: new advances and recent results, Curr. Opin. Struct. Biol. 8 (1998) 595–600.
[37] F. Tama, W. Wriggers, C.L. Brooks, Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory, J. Mol. Biol. 321 (2002) 297–305.
[38] M.M. Tirion, Large amplitude elastic motions in proteins from a single-parameter atomic analysis, Phys. Rev. Lett. 77 (1996) 1905–1908.
[39] R.C. Wade, H. Bohr, P.G. Wolynes, Prediction of water binding sites on proteins by neural networks, J. Am. Chem. Soc. 114 (1992) 8284–8285.
[40] T.P. Westcott, I. Tobias, W.K. Olson, Elasticity theory and numerical analysis of DNA supercoiling: an application to DNA looping, J. Phys. Chem. 99 (1995) 17926–17935.
[41] W. Wriggers, R.K. Agrawal, D.L. Drew, J.A. McCammon, J. Frank, Domain motions of EF-G bound to the 70S ribosome: insights from a hand-shaking between multi-resolution structures, Biophys. J. 79 (2000) 1670–1678.
[42] W. Wriggers, S. Birmanns, Using Situs for flexible and rigid-body fitting of multi-resolution single molecule data, J. Struct. Biol. 133 (2001) 193–202.
[43] W. Wriggers, P. Chacón, Modeling tricks and fitting techniques for multi-resolution structures, Structure 9 (2001) 779–788.
[44] W. Wriggers, P. Chacón, Using Situs for the registration of protein structures with low-resolution bead models from X-ray solution scattering, J. Appl. Crystallogr. 34 (2001) 773–776.
[45] W. Wriggers, R.A. Milligan, J.A. McCammon, Situs: a package for docking crystal structures into low-resolution maps from electron microscopy, J. Struct. Biol. 125 (1999) 185–195.
[46] W. Wriggers, R.A. Milligan, K. Schulten, J.A. McCammon, Self-organizing neural networks bridge the biomolecular resolution gap, J. Mol. Biol. 284 (1998) 1247–1254.
[47] W. Wriggers, K. Schulten, Investigating a back door mechanism of actin phosphate release by steered molecular dynamics, Proteins: Struct. Funct. Genet. 35 (1999) 262–273.
[48] Y. Yang, I. Tobias, W.K. Olson, Finite element analysis of DNA supercoiling, J. Chem. Phys. 98 (1992) 1673–1686.

Willy Wriggers is director of the Structural Bioinformatics program at University of Texas-Houston (UTH), where he is affiliated with the School of Health Information Sciences and the Institute of Molecular Medicine for the Prevention of Human Diseases. Wriggers received his Ph.D. degree in physics from the University of Illinois at Urbana Champaign, in 1998. Before joining the UTH faculty in 2003, Wriggers was assistant professor of molecular biology at The Scripps Research Institute (TSRI) in La Jolla, CA. In 2003 Wriggers was named Alfred P. Sloan Fellow in Computational and Evolutionary Molecular Biology. His interests include software development, biomolecular modeling, artificial neural networks, visualization, and biophysics.

Pablo Chacón received his Ph.D. degree in Biochemistry from the Universidad Complutense de Madrid, Spain, in 1999. From 2000 to 2003 he was research associate at TSRI in the laboratory of Willy Wriggers, where he developed software for image processing and biomolecular docking. He is currently research associate at the Centro de Investigaciones Biolgicas, Consejo de Investigaciones Cientificas in Madrid. His work interests include artificial neural networks and genetic algorithms and their applications to biophysical problems.

Julio A. Kovacs received his Ph.D. degree in mathematics from The Johns Hopkins University, in 1995. Afterwards he worked as a consultant for the National Space Activities Organization in Argentina and as assistant professor, both at the J.F. Kennedy University and at the National Technological University (Argentina). From 2001 to 2003 he was research associate at TSRI in the laboratory of Willy Wriggers. His research interests include numerical partial differential equations, general relativity, control systems, robotics, and mathematical biology. At TSRI he is currently developing methods for fast protein–protein docking.

Florence Tama received her Diplôme d' Étude Approfondie in Biophysics at Université Paul Sabatier (UPS, France), in 1997. During her doctoral studies, she collaborated with Yves-Henri Sanejouand at UPS and Nobuhiro Go at Kyoto University (Japan), to study the dynamical properties of biological systems with theoretical methods. During this time she developed a memory-efficient computational strategy for the block-based factorization of large Hessian matrices that arise in normal mode analysis (NMA) of biomolecular structures. Following her Doctorat in 2000 she continued her research in NMA at TSRI in the group of Charles L. Brooks III and in collaboration with the group of Willy Wriggers. Her current work focuses on extending NMA to large scale biomolecular machines using elastic networks at a reduced level of detail.

Stefan Birmanns received his Ph.D. in Computer Science at the University of Aachen, Germany, in 2003. Before graduation, from 1999 to 2003, he was as a doctoral candidate at Research Centre Jülich, John von Neumann Institute for Computing, Germany. In 2003 he joined the group of Willy Wriggers as a research associate, moving from TSRI to UTH. He currently designs and administrates a virtual reality laboratory and supervises students involved in the graphics development in Germany and Houston. His current research interests are haptic rendering, virtual-reality and visualization methods and the application of these methods to biophysical problems.