PEPSI-Dock: a detailed data-driven protein–protein interaction potential accelerated by polar Fourier correlation

Emilie Neveu, David Ritchie, Petr Popov, Sergei Grudinin

Bioinformatics, 32, 2016, i693–i701
doi: 10.1093/bioinformatics/btw443
ECCB 2016

PEPSI-Dock: a detailed data-driven protein–protein interaction potential accelerated by polar Fourier correlation

Emilie Neveu[1], David W. Ritchie[2], Petr Popov[1,3] and Sergei Grudinin[1,*]

[1]Inria/University Grenoble Alpes/LJK-CNRS, F-38000 Grenoble, France, [2]Inria Nancy – Grand Est, 54600 Villers-les-Nancy, France and [3]Moscow Institute of Physics and Technology, Dolgoprudniy, Russia

*To whom correspondence should be addressed

Abstract

Motivation: Docking prediction algorithms aim to find the native conformation of a complex of proteins from knowledge of their unbound structures. They rely on a combination of sampling and scoring methods, adapted to different scales. Polynomial Expansion of Protein Structures and Interactions for Docking (PEPSI-Dock) improves the accuracy of the first stage of the docking pipeline, which will sharpen up the final predictions. Indeed, PEPSI-Dock benefits from the precision of a very detailed data-driven model of the binding free energy used with a global and exhaustive rigid-body search space. As well as being accurate, our computations are among the fastest by virtue of the sparse representation of the pre-computed potentials and FFT-accelerated sampling techniques. Overall, this is the first demonstration of a FFT-accelerated docking method coupled with an arbitrary-shaped distance-dependent interaction potential.

Results: First, we present a novel learning process to compute data-driven distant-dependent pairwise potentials, adapted from our previous method used for rescoring of putative protein–protein binding poses. The potential coefficients are learned by combining machine-learning techniques with physically interpretable descriptors. Then, we describe the integration of the deduced potentials into a FFT-accelerated spherical sampling provided by the Hex library. Overall, on a training set of 163 heterodimers, PEPSI-Dock achieves a success rate of 91% mid-quality predictions in the top-10 solutions. On a subset of the protein docking benchmark v5, it achieves 44.4% mid-quality predictions in the top-10 solutions when starting from bound structures and 20.5% when starting from unbound structures. The method runs in 5–15 min on a modern laptop and can easily be extended to other types of interactions.

Availability and Implementation: https://team.inria.fr/nano-d/software/PEPSI-Dock.

Contact: sergei.grudinin@inria.fr

1 Introduction

Protein–protein interactions (PPIs) can provide useful targets for the development of novel antiviral and antibacterial therapeutic agents. PPIs are also of particular interest for cancer treatments where there is a growing need to target more precisely unhealthy cells and to avoid unnecessary damage to healthy cells. However, such approaches require knowledge of the three-dimensional (3D) structure of the target complex. Exploiting existing protein structures in the Protein Data Bank (PDB) can help, but only a small fraction of these structures are complexes (Berman *et al.*, 2000). Therefore, computational protein docking predictions, being low-cost and easy to perform, would be very attractive if they could reliably identify the most probable binding conformation of a pair of interacting proteins.

Currently, many algorithms can produce high quality predictions of molecular complexes (Janin, 2005; Méndez *et al.*, 2003), but identifying the correct binding mode from amongst those predictions remains a difficult problem, especially when the proteins are large or when they exhibit large conformational changes on binding (Bonvin, 2006). Generally, existing docking techniques use combinations of methods to cover the conformational search space and to score each candidate conformation. They often decompose the docking problem into stages using different degrees of accuracy. A common approach is to start with a simple scoring model describing the large scale movements of the proteins in order to restrict the 3D global space to a few promising binding sites that will be further explored with a more precise scoring model which takes into account smaller scale effects such as steric clashes.

The work we present here focuses on improving the first stage of the docking pipeline. We believe it will also improve the subsequent pose refinement calculation because important 3D solutions may be missed if the low resolution model is too crude. We therefore developed an original learning process that combines robust machine learning techniques with physically interpretable descriptors and we adapted this to the fast Fourier transform (FFT) accelerated spherical sampling provided by the Hex library (Ritchie *et al.*, 2008). Thus the present work is the first attempt to introduce an arbitrary interaction potential into a FFT-engine.

While it is not possible to list all the docking algorithms here, we will give a brief overview of the existing sampling and scoring methods that share the same objectives as our approach. Such a variety of methods exists because predicting the structures of a complex of proteins is not only about how well two shapes will fit together. Proteins are chains of amino acids with more or less freedom in their movements. Placed in a solvent, they may bind to each other to reach a more stable arrangement of atoms that corresponds to the state where the difference of the Gibbs free energy upon binding is minimum. This state depends on physical processes such as electrostatic interactions, hydrogen bonds, side chain and backbone flexibility, rearrangements of solvent molecules and entropic interactions between flexible proteins and their environment. In theory, predicting the stable conformation of the complex can be performed by running molecular dynamics (MD) simulations, where atomic trajectories are subject to a mechanical force field that aims to describe the potential energy of the system through Newton's equations of motion. (Brooks *et al.*, 1983; Case *et al.*, 2005; Hess *et al.*, 2008). Thus, MD methods can be referred to as physics-based methods.

Currently, extending MD force-field models with quantum mechanical parameters and better solvation models is an active field of research (Kuhn *et al.*, 2005; Chaskar *et al.*, 2014). In any case, whether they rely on strong physics theory or not, such computations imply many unknowns to find from ill-conditioned systems of equations and parameters. Typically, MD-based approaches cannot be used at the first stage of the docking pipeline. Instead, more practical approaches have been developed which have both advantages and disadvantages in the way that they deal with the challenge of combining physical laws with observed data.

Setting aside the above physics-based methods, many scoring methods are hybrids which may be categorized into three groups, namely empirical scoring functions, data-driven potentials and descriptor-based scoring functions (Liu and Wang, 2015). Empirical scoring functions consist of a linear combination of terms that reflect important factors such as hydrophobic contacts, hydrogen bonding, accessible and buried surface area (e.g. Böhm, 1994; Eldridge *et al.*, 1997; Friesner *et al.*, 2004; Wang *et al.*, 2002). Multivariate regression analysis is used to determine the weights of the different terms such that scoring provides good agreement with a training set of complexes. These scoring methods can be easily adapted to a specific type of interaction (Korb *et al.*, 2009). However, they strongly depend on the dataset.

Data-driven potentials are typically a sum of pairwise statistical potentials calculated from structural information from databases of molecular complexes (Chuang *et al.*, 2008; Huang and Zou, 2011; Mooij and Verdonk, 2005; Muegge and Martin, 1999; Popov and Grudinin, 2015; Zhou and Skolnick, 2011). These methods rely on the assumption that interactions are more frequently observed if they induce stability and that native molecular complexes will possess distinct structural features compared to non-native structures. Data-driven potentials are often derived from the inverse Boltzmann statistics with respect to a known reference state (Koppensteiner and Sippl, 1998), and can be combined with entropic terms (Gohlke *et al.*, 2000; Huang and Zou, 2010). Scoring with these potentials is generally more computationally efficient. Hence, data-driven potentials are very suitable for fast docking and virtual screening protocols.

On the other hand, descriptor-based methods use machine learning techniques such as neural networks, random forest or support vector machines (SVMs) to learn from a large set of data (Kinnings *et al.*, 2011; Li *et al.*, 2013; Zilian and Sotriffer, 2013). The training process provides a relationship between the observed data and the molecular descriptors that can be of any nature, e.g. geometrical, topological or electronic structure-based. However, the use of descriptor-based methods raises some important questions. First, the reasons for selecting a certain combination of descriptors are generally not clearly specified. Second, the not necessarily linear relationship between the descriptors is not physically interpretable (Gabel *et al.*, 2014).

In many docking methods, scoring and sampling are closely linked because the choice of scoring function strongly affects the shape of the energy landscape to explore. In general, sampling algorithms can be grouped into three categories, exhaustive search, deterministic search and random sampling search methods. An exhaustive search spans the entire discretized search space and computes the energy at each point—the finer the discretization, the more it is precise. Obviously, exhaustive search algorithms are the most robust algorithms, but their cost is typically very high, especially in high-dimensional search spaces. Indeed, the cost increases exponentially with the number of degrees of freedoms. To reduce the computational cost of exhaustive sampling, FFT-based techniques are used (Katchalski-Katzir *et al.*, 1992; Kozakov *et al.*, 2006; Ritchie *et al.*, 2008). They take advantage of the fast computations of correlations in Fourier space.

Deterministic search algorithms rely on optimization theory to find extremes in the energy landscape using knowledge of its first and second derivatives. These methods are very powerful and efficient when the problem can be defined as a convex optimization (Nocedal and Wright, 2000) where only one solution exists. Unfortunately, in blind docking searches, the optimization problem is strongly non-convex. Thus, deterministic algorithms are generally only used in the refinement stage of a docking protocol. Instead, random sampling algorithms such as simulated annealing or Monte Carlo searches are used (Gray *et al.*, 2003). These stochastic methods are less likely to converge to local minima, and their convergence towards a global minimum is not guaranteed.

Here, we aim to improve the first stage of the docking pipeline by performing FFT-accelerated exhaustive search calculations with an accurate approximation of the binding free energy expressed as a linear combination of distance dependent potentials. For this purpose, we developed the PEPSI-Dock method, where PEPSI stands for Polynomial Expansion of Protein Structures and Interactions. PEPSI-Dock predicts 3D PPIs using a simple to compute but detailed expression of physical laws in conjunction with a very fast FFT-accelerated exhaustive search. More precisely, we adapted the computation of the data-driven potentials from our previous work (Grudinin *et al.*, 2015; Popov and Grudinin, 2015) to the 3D Spherical Polar Fourier basis. This allows us to use spherical polar Fourier transforms, as implemented in Hex (Ritchie and Venkatraman, 2010), to rapidly compute energy overlap integrals during the exhaustive search for putative docking poses.

The main novelty of the work presented here consists of the application of an arbitrary-shaped data-driven potential to an exhaustive FFT search algorithm and the validation of its efficiency for blind docking predictions. The PESPI-Dock scoring function is mainly inspired by the SVM technique, but it combines physically interpretable descriptors. We deduce the potentials using a convex optimization of a well-defined problem. Our potentials represent 210 types of distance-dependent pairwise interactions of free functional form. We optimize the shapes of the potentials such that the ability of the scoring function to discriminate native from non-native poses is maximized. While our potential still depends on the dataset, its expression is simpler than a physics-based potential and does not depend on a reference state. Once learned, the potentials can be stored and used later for docking predictions, almost with no cost. The only current limitation is the linearity with respect to coordinates of atoms and the rigid-body assumption.

This approach is very versatile and can be readily adjusted to other types of interactions. Indeed, we have already used the same methodology to derive potentials for the refinement step of protein–ligand docking (Grudinin *et al.*, 2015) and *ab initio* protein–protein docking (Popov and Grudinin, 2015) (since we did not use any homology knowledge or MD refinement). The latter proved to be very successful in a recent CAPRI/CASP docking assessment exercise (Lensink *et al.*, 2016), where our methodology was ranked sixth overall and the top template-free *ab initio* approach.

2 Approach

Figure 1 illustrates the PEPSI-Dock method, which consists of two main stages. The first stage is the learning phase (steps 1–4 in Fig. 1), where starting from a set of native and non-native complexes, we deduce 210 atomic one-dimensional (1D) distance-dependent interaction potentials. These potentials are based on commonly used physical assumptions accounting for 210 different types of interactions that depend on the atom characteristics (chemical element, aromaticity, polarity, etc.). They are constructed with the specific objective of discriminating native structures from the non-native ones, thanks to the use of a rigorous learning optimization algorithm inspired by the SVM method. The feature space for the optimization problem consists of the geometrical and chemical properties extracted from interfaces between proteins. The second stage is the docking prediction phase (steps 5–7 in Fig. 1), where starting from a pair of proteins, a list of putative binding poses with low binding free energy conformations is computed. The possible conformations are explored using a global exhaustive search algorithm. It represents the protein geometry through 20 3D Gauss-Laguerre expansion coefficients and uses the Spherical Polar Fourier transform to rapidly compute the energy overlap integrals. The main contribution of the paper is the combination of these two stages. More exactly, we combine the arbitrary-shaped data-driven potentials with a FFT-exhaustive search algorithm. To do so, we define the potentials in the Gauss-Laguerre basis. Thanks to the spherical symmetry of an individual potential, we analytically convert 1D pairwise contributions to 3D potential maps expressed in the Gauss-Laguerre basis (see Fig. 3).

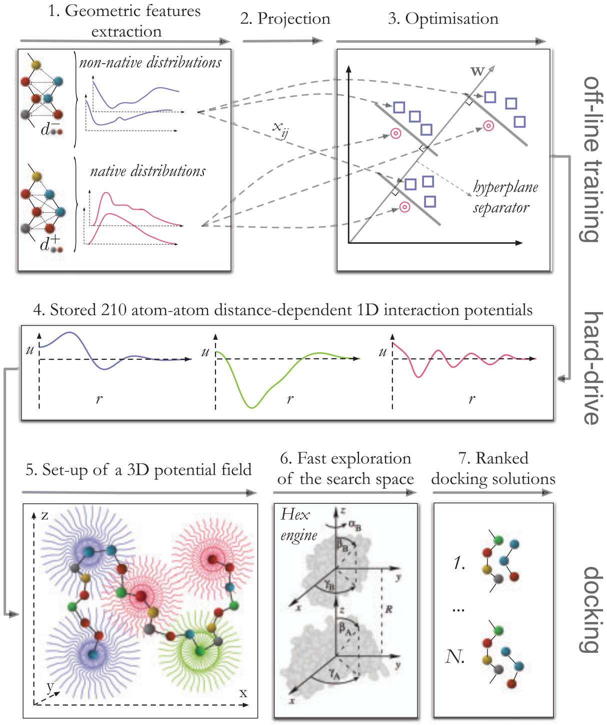

Fig. 1. Schematic representation of the PEPSI-Dock method. The first stage is learning the interaction potentials from the geometry and chemistry of protein–protein interfaces. The second, docking stage, uses the 1D distance-dependent interaction potentials to exhaustively score a list of generated conformations, which are produced by the *Hex* library on a 6D spherical grid

3 Methods

3.1 Learning from known interfaces

The PEPSI-Dock scoring ranks the docking predictions using 1D atomistic distance-dependent data-driven interaction potentials. We distinguish the atoms according to their types. These types are specific for protein–protein interactions and take into account chemical properties of the corresponding atoms such as their chemical element, aromaticity, polarity and hybridization. We adapted these types from Huang and Zou (2008), having in total 20 different atom types. Due to the symmetry of the interactions, we thus need to define 210 (i.e. 20 × 21/2) different interaction potentials. Each of these potentials approximates atom–atom pairwise interaction as a function of the distance between the atoms. The first step of the learning process is the extraction of geometric features from the structural data of native protein complexes and their generated non-native structures (decoys). These features are encoded using the number density of the interactions of a certain type at a certain distance (step 1 in Fig. 1). Then, the densities are projected onto a truncated Laguerre polynomial basis (step 2 in Fig. 1) and finally, an optimization algorithm computes the potentials such that the derived scoring function can correctly discriminate the native conformations from the non-native ones (step 3 in Fig. 1).

3.1.1 Physical model

Let us consider P native protein–protein complexes C_{i0}, $i = [1, P]$. For each native configuration, we generate D non-native configurations (decoys) by applying rigid transformations to the smaller molecule (the ligand) and obtain C_{ij} with $j \in [1, D]$, where the first index indicates the protein complex and the second index indicates the generated decoys. Thus, for each complex, we have $D+1$ conformations, 1 native and D non-native. Our aim is to find a *scoring functional* E such that the following inequalities hold,

$$E(C_{i0}) < E(C_{ij}), \quad \forall i \in [1, P], \quad \forall j \in [1, D] \qquad (1)$$

In order to solve this problem, we need to make some assumptions. First, we represent the proteins as a set of discrete interaction sites located at the centres of the atomic nuclei. Each interaction site is defined according to the properties of its atom, such as the chemical element, its aromaticity, hybridization state and polarity. Thus, an interaction between two sites can be regarded as an interaction between two types of atoms. This results in a total of $M \times (M+1)/2$ pairs of interactions, with M being the total number of interaction sites. Second, we assume that E depends only on the distribution of the distances between two interaction sites, i.e. the number of site pairs at a certain distance, one site being located on the receptor (the larger protein) and the other on the ligand (the smaller protein),

$$E(C_{ij}) = E(n_{ij}^{11}(r), .., n_{ij}^{kl}(r), .., n_{ij}^{MM}(r)) = E(n_{ij}(r)), \tag{2}$$

where, for each decoy, C_{ij}, $n_{ij}^{kl}(r)$ is the *number density of site-site pairs* at a distance r between two sites k and l, with site k located on the receptor and site l located on the ligand. Third, we assume that interactions are short-range, and may be neglected if the distance between two interaction sites is larger than a certain cutoff distance $r_{\max}$. This allows us to restrict the information extracted from the complexes to their interfaces. We use a cutoff distance value of 10 Å, which has been widely used in previous approaches (Chae *et al.*, 2010;Chuang *et al.*, 2008; Huang and Zou, 2008; Maiorov and Grippen, 1992; Qiu and Elber, 2005; Tobi and Bahar, 2006), and which gave good results in our earlier experiments (Grudinin *et al.*, 2015). Finally, we assume that E is a linear functional,

$$E(n_{ij}(r)) = \sum_{k=1}^{M} \sum_{l=k}^{M} \int_{0}^{r_{\max}} n^{kl}(r) f^{kl}(r)\ dr, \tag{3}$$

where $f^{kl}(r)$ are the unknown *interaction potentials* that we need to determine. In order to determine unknown potentials $f^{kl}(r)$, we decompose them along with the number densities $n^{kl}(r)$ in a Laguerre polynomial basis,

$$\begin{aligned} f^{kl}(r) &= \sum_{q=0}^{\infty} w_q^{kl} \psi_q(r), \qquad r \in [0; \infty] \\ n^{kl}(r) &= \sum_{q=0}^{\infty} x_q^{kl} \psi_q(r), \qquad r \in [0; \infty], \end{aligned} \tag{4}$$

where $\psi_q(r)$ are the Laguerre basis functions orthogonal on $[0; \infty]$, and w_q^{kl} and x_q^{kl} are the expansion coefficients of $f^{kl}(r)$ and $n^{kl}(r)$, respectively.

To have a sparse representation, the scoring functional E can be truncated up to the order Q as

$$E(n_{ij}(r)) \approx \sum_{k=1}^{M} \sum_{l=k}^{M} \sum_{q=0}^{Q} w_q^{kl} x_q^{kl} = (\mathbf{w} \cdot \mathbf{x}), \ \ \mathbf{w}, \mathbf{x} \in \mathrm{R}^{Q \times M \times (M+1)/2} \tag{5}$$

Looking again at Figure 1, we can see that $\mathbf{x}$ represents the expansion coefficients of the sum of *number densities of site–site pairs* observed at the interface of the decoy C_{ij}. These are our geometric features, whereas $\mathbf{w}$ is the set of expansion coefficients of the learned interaction potentials, computed solving an optimization problem. We will refer to the vector $\mathbf{w}$ as to the *scoring vector*, whose value is to be determined, and to the vector $\mathbf{x}$ as to the *structure vector* that is computed from the structural data.

3.1.2 Optimization algorithm

Here we briefly explain the optimization algorithm used to determine the *scoring vector* $\mathbf{w}$. This will be used to define a scoring function that discriminates native protein–protein interfaces from non-native ones. We will start by defining a rigorous optimization algorithm inspired by the SVM method that interprets each complex C_{ij} as a structure vector $\mathbf{x}_{ij}$, with the corresponding label y_{ij}, which indicates whether the complex is native or not. More precisely, Equation (5) allows us to rewrite the set of inequalities (1) as a *convex optimization problem*, which consists in finding $\mathbf{w}$ that minimizes the empirical risk with a regularization penalty (Scholkopf and Smola, 2001) as

$$\min_{\mathbf{w}} \ \frac{\lambda}{2} \Omega(\mathbf{w}) + L(X; \mathbf{w}), \tag{6}$$

where $X = (\mathbf{x}_{ij}, y_{ij})$ is the training set of labeled structure vectors, $L(X; \mathbf{w})$ is the empirical risk function, and $\Omega(\mathbf{w})$ is the regularization penalty, which aims to prevent over-fitting and compensates the lack of statistics for rare events. We use the log-loss function to describe the empirical risk as a smooth approximation to the hinge-loss function and the two-norm for the regularization (Popov and Grudinin, 2015). We should mention that parameter λ is the only adjustable parameter in our model that determines the importance of the regularization term with respect to the empirical risk. Its value is determined using a cross validation procedure such that the corresponding potential gives the best prediction on complexes that have not been used during the learning process.

We solve Equation (6) using a quasi-Newton optimization in primal (Boyd and Vandenberghe, 2004). This is a general and robust convex optimization procedure that has been proved to converge towards the optimum independently of the starting configuration. Since our problem is convex by construction, this guarantees the existence of a single minimum (Boyd and Vandenberghe, 2004; Nocedal and Wright, 2000).

3.2 Predicting unknown structures

To be able to apply FFT acceleration of the energy calculation of different conformations, we assume E to be a linear sum of overlap integrals of the receptor potentials $f(\mathbf{x})$ with ligand densities $g(\mathbf{x})$,

$$E = \sum_{\text{atom type}\, k} \iiint_V \sum_{L_k} g(\mathbf{x} - \mathbf{x}_{L_k}) \sum_{\text{atom type}\, l} \sum_{R_l} f^{kl}(\mathbf{x} - \mathbf{x}_{R_l}) dV, \tag{7}$$

where indices k and l run from 1 to the number of atom types M. R_l runs over the receptor atoms of type l, and L_k over the ligand atoms of type k. With M equal to 20, the scoring function E is a sum of 20 different 3D overlap integrals. We define the density of ligand atoms $\sum_{L_k} g(\mathbf{x} - \mathbf{x}_{L_k})$ to be a sum of spherical Gaussians centered at each atom with coordinates of $\mathbf{x}_{L_k}$. Finally, $f^{kl}(\mathbf{x})$ is a 3D spherically symmetric interaction potential. Thus, $\sum_{R_l} f^{kl}(\mathbf{x} - \mathbf{x}_{R_l})$ is a sum of pairwise spherically symmetric interaction potentials centered at receptor atoms of coordinates $\mathbf{x}_{R_l}$. From now on, we will use f^k to refer to $\sum_l \sum_{R_l} f^{kl}(\mathbf{x} - \mathbf{x}_{R_l})$, and g^k to refer to $\sum_{L_k} g(\mathbf{x} - \mathbf{x}_{L_k})$. Thus, the scoring energy E is given as a sum of 20 overlap integrals of the densities g^k, with potentials f^k. For practical purposes, we represent both functions g^k and f^k as spherical grids in the Laguerre polynomial basis.

Knowing the general form of the scoring function E, we now refer to the lower part of Figure 1 to describe the different steps of the docking process. Below, we will first explain the fast exploration of the docking search space that makes reference to the *Hex* sampling algorithm and its efficient way to compute the overlap integrals between functions g^k and f^k.

3.2.1 Exploration of the search space

Here we briefly describe the main components of the conformational space exploration algorithm. For more detail, we refer the reader to the reference paper (Ritchie and Kemp, 2000), which introduces the *Hex* docking method. The sampling algorithm defines the potentials $f(\mathbf{x})$ and the densities $g(\mathbf{x})$ attached to the proteins in spherical coordinates (r, θ, ϕ). The search space is then explored by computing the energies of each generated conformation using a discretized set of values for the six degrees of freedoms. These are the distance between the proteins R and the five Euler angles, two receptor angles (β_A, γ_A) and three ligand angles $(\alpha_B, \beta_B, \gamma_B)$.

Figure 2 schematically shows the definition of the search space. Each 'docking' conformation expresses rigid transformations of the ligand and the receptor molecules. Thus, the scoring energy defined in Equation (7) should be rewritten as a function of these 6 degrees of freedom,

$$\begin{aligned} E &= E(R, \beta_A, \gamma_A, \beta_B, \gamma_B, \alpha_B) \\ &= \sum_k \iiint_V \left(\widehat{R}(0, \beta_A, \gamma_A) f^k(r, \theta, \phi)\right) \\ &\quad \times \left(\widehat{T}_z(R)\widehat{R}(\alpha_B, \beta_B, \gamma_B) g^k(r, \theta, \phi)\right) dV. \end{aligned} \tag{8}$$

Here we assume that both proteins are initially located at the origin, and $\widehat{T}$ and $\widehat{R}$ are the translation and rotation operators, respectively.

In order to efficiently compute integrals in Equation (8), we use truncated polynomial expansions of the functions f^k and g^k in a spherical polar Fourier (SPF) basis such that

$$\iiint_V f^k(r, \theta, \phi) \,.\, g^k(r, \theta, \phi) dV = \sum_{nlm}^{N} \mathbf{f}^k_{nlm} \mathbf{g}^{*k}_{nlm}, \tag{9}$$

where N is the expansion order, and the star sign denotes complex conjugated coefficients. We should emphasize that the SPF representation of the functions f^k and g^k allows for a very fast exploration of the rotational degrees of freedom (DOFs) during the docking search. It is also interesting to note that spherically symmetric functions have a sparse representation in this basis. More precisely, spherically symmetric functions *f*(r) centered at the origin have only N non-zero expansion coefficients $\mathbf{f}_{n00}$.

The SPF expansions can be seen as the extension of the Fourier expansions into spherical coordinates. The SPF basis functions are composed of spherical harmonics multiplied by a radial orthogonal polynomial, which are the Gaussian-type orbitals in our case (Ritchie and Kemp, 2000). The SPF basis functions form an orthonormal basis in 3D and well encode a certain spherical volume, whose size depends on the expansion order N and on the scaling of the radial basis functions. Both rotation and translation operators have an analytic representation in the SPF basis (Ritchie, 2005). Rotations are particularly fast to compute, because each group of spherical harmonics of the same order transforms among themselves using the Wigner rotation matrix.

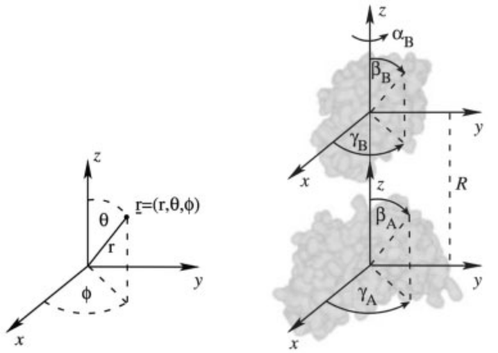

Fig. 2. Left: the relationship between the spherical polar (r, θ, ϕ) and Cartesian (*x*, *y*, *z*) coordinate systems; right: schematic illustration of the 6D rigid body search space in terms of one translational coordinate, *R*, and five Euler rotational coordinates, (β_A, γ_A) and $(\alpha_B, \beta_B, \gamma_B)$, assigned to the receptor and ligand, respectively. Following the usual ZYZ Euler angle convention, β rotations refer to the y−axis, while α and γ rotations refer to the z−axis. Figure reproduced from Ritchie *et al.* (2008) with permission from Oxford University Press

The Hex sampling algorithm can evaluate energies from Equation (8) at different values of the angular degrees of freedom simultaneously using the multidimensional correlation theorem. This is implemented with either 3D or 5D FFTs and allows to accelerate the search by up to two orders of magnitude (Ritchie *et al.*, 2008). The correlation procedure is then repeated for each value of the translation degree of freedom.

At the last step, after the computation of the scoring energy E for all the poses, we spatially cluster those that have the least energy. We do it thanks to a simple clustering algorithm based on a constant-time RigidRMSD library (Popov and Grudinin, 2014). This allows to cluster together conformations have relative root mean squared (RMS) deviations (RMSDs) below a certain threshold.

3.2.2 Shape of f ^{kl} and g^{kl} functions

First, we will demonstrate that for any spherically symmetric functions $g(\mathbf{x} - \mathbf{x}_{L_k})$ and $f(\mathbf{x} - \mathbf{x}_{L_k})$, the 3D expression of the scoring energy in Equation (7) can be reduced to a linear combination of 1D integrals. Indeed, regrouping the summations, splitting the 3D integral into the radial and the angular part, and taking the angular integrals analytically, it is easy to obtain

$$E = \sum_{k,l=1}^{M} \sum_{R_l} \int_0^\infty f^{kl}(r) \sum_{L_k} g_{|1D}(r - ||\mathbf{x}_{L_k} - \mathbf{x}_{R_l}||_2) dr, \tag{10}$$

where $g_{|1D}(r)$ is a 1D radial projection of the function $g(\mathbf{x})$,

$$g_{|1D}(r) = \int_0^{2\pi} \int_0^{\pi} g(r, \theta, \phi)\, r^2 \sin\theta \, d\theta \, d\phi \tag{11}$$

This scoring energy representation has the *same functional form* as our previous expression for the binding free energy in Equation (3). However, to use this result, we need to define the precise shape of the functions $g(\mathbf{x} - \mathbf{x}_{L_k})$ and their 1D radial projections. More specifically, we define each individual ligand density function $g(\mathbf{x} - \mathbf{x}_{L_k})$ in Equation (7) as a normalized 3D spherical Gaussian centered at $\mathbf{x}_{L_k}$,

$$g(\mathbf{x} - \mathbf{x}_{L_k}) = \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} e^{-\frac{(\mathbf{x} - \mathbf{x}_{L_k})^2}{2\sigma^2}}, \tag{12}$$

where σ is a width parameter constant for all atoms. Then, we can compute its 1D radial projection as

$$g_{|1D}(r - r_{L_k}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r - r_{L_k})^2}{2\sigma^2}} \frac{r}{r_{L_k}} \left(1 - e^{-2\frac{r r_{L_k}}{\sigma^2}}\right), \tag{13}$$

where $r_{L_k} = ||\mathbf{x}_{L_k}||_2$. Thus, we have demonstrated that if the number densities $n_{kl}(r)$ in Equation (3) are computed as a sum of radial Gaussian projections,

$$n^{kl}(r) = \sum_{L_k} g_{|1D}(r - ||\mathbf{x}_{L_k} - \mathbf{x}_{R_l}||_2), \tag{14}$$

then the scoring potentials $f^{kl}(r)$ obtained through the convex optimization procedure Equation (6) will be valid for the evaluation of

3D overlap integrals with spherical Gaussians defined in Equation (12).

Finally, for practical reasons, we decompose both $n^{kl}(r)$ and $f^{kl}(r)$ functions onto a Laguerre polynomial basis, as given by Equation (3.1.1). We do so, so that the learned scoring vector $\mathbf{w}$ is a set of expansion coefficients $\mathbf{f}^{kl}_{n00}$ from Equation (9) for each interaction type pair kl. The structure vector $\mathbf{x}_{ij}$, as we mentioned above, is a set of Laguerre expansion coefficients of the number densities $n^{kl}_{ij}(r)$ computed with Equation (14) for the interfaces of decoys C_{ij}. We should mention that thanks to the spherical symmetry of the interaction potentials, there are only N non-zero coefficients for each interaction type pair kl, $\mathbf{f}^{kl}_{n00}$. These are obtained offline via the optimization procedure Equation (6), stored on the hard-drive and loaded in the PEPSI-Dock at the beginning of the docking process.

At the pre-processing stage of our docking method (see step 5 in Fig. 1), we position all the interaction potentials described with their expansion coefficients to the centres of atomic nuclei. To do this, we first centre each of spherically symmetric coefficients $\mathbf{f}^{kl}_{n00}$ at the origin. Then, we analytically translate these along the z-axis as it is explained in Ritchie (2005) to obtained axial-symmetric coefficients $\mathbf{f}^{kl}_{nl0}$. After, we rotate $\mathbf{f}^{kl}_{nl0}$ with two consecutive rotation operators, one about the y-axis by angle θ, and another about the z-axis by angle ϕ, to obtain the final expansion coefficients $\mathbf{f}^{kl}_{nlm}$. Figure 3 summarizes these steps.

3.3 Datasets and experimental settings

3.3.1 Training *set*

We used the training database of 851 non-redundant protein–protein complex structures prepared by Huang and Zou (2008). This database contains protein–protein complexes extracted from the PDB (Berman *et al.*, 2000) and includes 655 homodimers and 195 heterodimers. We used only the structures of the heterodimers for the training.

For each native conformation of the complex from the training set, we added 199 non-native conformations (decoys) generated by applying rigid transformations to the ligand. These are the false-positive predictions sorted by the score and generated by the default Hex shape-complementarity function. We defined the decoys as the structures with a ligand RMS (L-RMS) greater than 10 Å from the native pose. L-RMS is the RMS deviation between the true and the predicted ligands when the receptors are aligned. The final training set is composed of 195 native complex structures with 38 805 decoys, resulting in a total of 39 000 complex structures.

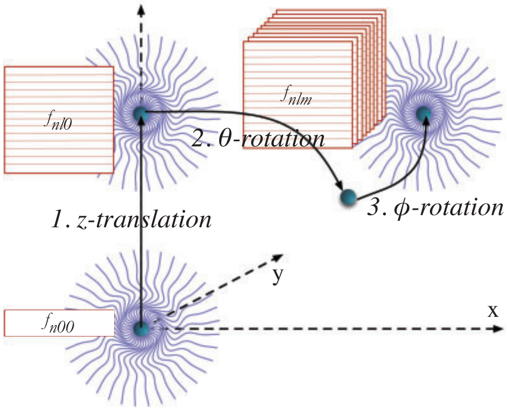

Fig. 3. Summary of PEPSI-Dock 3D potential set-up. We start with a vector of 1D coefficients centered at the origin. Then, we translate these along the z-axis to obtain a matrix of 2D coefficients. Finally, we rotate this matrix with two consecutive rotation operators to obtain the final tensor of 3D coefficients. This operation is repeated for each atom of the receptor and the ligand

3.3.2 Test set

For the test set, we used the Protein Docking Benchmark, which is the most widely used set of protein complexes collected both in the bound and in the unbound forms. Version 5.0 of the benchmark was released recently (Vreven *et al.*, 2015) and consists of 230 protein complexes. The complexes are categorized based on the structural differences between the bound and the unbound forms of the binding partners, accounting for the fraction of non-native residues contacts and the interface RMS (I-RMS) deviation. This is the RMS deviation of C_α atoms of interface residues calculated after finding the best superposition of bound and unbound interfaces. For example, complexes are regarded as ‘difficult’ if their I-RMS is greater or equal to 2.2 Å. We used the bound and the unbound sets from this benchmark to assess the quality of PEPSI-Dock predictions.

3.3.3 Feature extraction

We extracted structural features from the interfaces in the training set using all pairs of atoms at a distance lower than the cutoff distance $r_{\max}$ of 10 Å. The standard deviation of the Gaussian densities was set to $\sigma = 1$ Å. We decomposed the interaction potentials and the number densities in the SPF basis using an expansion order of $N = 30$, with the support for the radial functions of 30 Å. We used 20 atom types defined by the classification of all heavy atoms in standard amino acids according to their element symbol, aromaticity, hybridization and polarity (Grudinin *et al.*, 2015). This leads to the total number of 210 types of interactions, which defines the size of the feature space of 6300. In this space, we solved the optimization problem Equation (6), with the only adjustable parameter λ, which defines the weight of the regularization term. Finally, to avoid over-fitting, we used a 2-fold cross validation procedure that determined the optimal value of the λ parameter. Figure 4 shows the deduced potentials for the interactions between guanidine nitrogen and carboxyl oxygen on the left and two aliphatic carbons on the right. We can see that the reconstructed potentials are very smooth and similar to our previous potentials learned in a different polynomial basis (Popov and Grudinin, 2015) and also similar to those obtained with different techniques (Huang and Zou, 2008).

3.3.4 Rigid-body docking parameters

We discretized the search space using the radial step size of 1 Å and the angular step size of 7.5°. We set the maximum radial separation distance to 40 Å. We clustered the solution with a distance threshold of 5 Å. When doing tests on bound structures of protein–protein complexes, we randomly perturbed the initial ligand position to exclude trivial rotational solutions from the results.

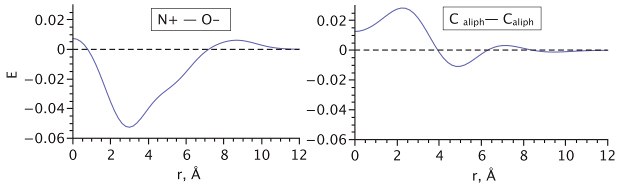

Fig. 4. Two selected data-driven potentials deduced from the training set using the convex optimization technique. Left: potential for the interactions of guanidine nitrogen with carboxyl oxygen. Right: potential for the interactions between two aliphatic carbons. The dashed line corresponds to $y = 0$

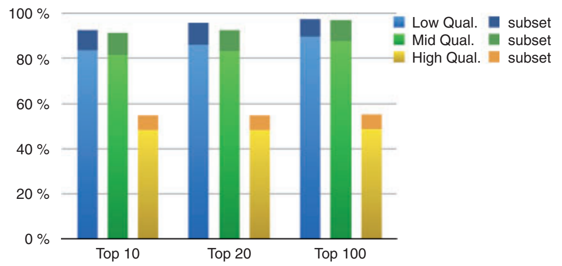

Fig. 5. Percentage of success rates of PEPSI-Dock on the 195 bound heterodimers from the Training Set using L-RMS criteria. With darker colours, results have been restricted to the 163 bound heterodimers of the Training Set for which the separation distance $\leq$ 30 Å

3.3.5 Quality assessment

For each tested complex, we clustered the docking poses and computed the L-RMS and I-RMS deviation between the native pose and the lowest energy candidate of each cluster. When handling unbound conformations, RMS deviations were computed after having superposed the unbound structures with bound ones. We define the prediction to be of a high quality if L-RMS $\leq$ 1 Å. Mid-quality corresponds to L-RMS $\leq$ 5 Å, and low-quality to L-RMS $\leq$ 10 Å. In case of I-RMS, threshold are adjusted to 1 Å (high qual.), 2 Å (mid) and 4 Å (low) following to CAPRI criteria. We define the success rate as the percentage of complexes for which a good quality prediction (high, mid or low) has been found among the lowest energy conformations.

3.4 Results and discussion

First, we validated the method on 195 protein complexes from the training set. We should emphasize that during the training stage, we used only 199 structures of decoys for each of the complex from this set, whereas PEPSI-Dock exhaustively explores and ranks about 10^9 putative binding poses. Moreover, using weights, the native and non-native poses are accounted for in an equal way. Thus, we believe that we do not have any significant bias or over-fitting in this test. Figure 5 presents the results on this test. The success rates are sufficiently high. In order to better understand the pitfalls of the method, in Figure 6, we plotted a histogram of the top-10 mid-quality success rates as a function of the distance separating the centres of masses between the protein and the ligand. We can clearly see that complexes with larger separation distances are more difficult to predict. This can be explained by two factors. First, the quality of our polynomial approximations degrades with the distance. Since the support of our radial functions is 30 Å, we should expect a rapid degradation of the prediction results beyond this distance. Second, the sampling method uses a spherical grid that makes the precision of the search decrease with the distance to the origin of the grid. On the other hand, Figure 6 compares the quality of prediction of PEPSI-Dock with the original method implemented in Hex with the shape complementarity plus electrostatics scoring function. We can see that PEPSI-Dock achieves significantly better results than Hex, especially for complexes with the separation distance less than 30 Å.

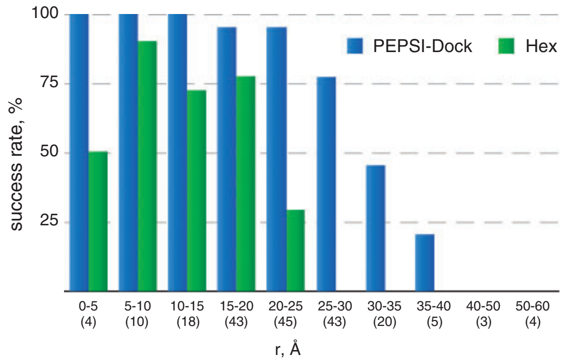

Fig. 6. The success rates of the PEPSI-Dock and Hex methods on the Training Set as a function of different separation distances between the receptor and the ligand proteins. PEPSI-Dock results are plotted in blue. Hex results are plotted in green. The success rate is defined as the percentage of structures that have at least one solution with L-RMS $\leq$ 5 Å found in the top 10 solutions. The separation distances are measured between the centres of mass of the two proteins in the native state. The number of complexes with the corresponding separation distance is shown inside the brackets

To better analyze the results, we now assess only the complexes with a separation distance less or equal to 30 Å. This includes 163 complexes from the Training Set. Figure 5 also presents the success rates for this subset with darker color. We achieve 91.4% instead of 81.5% for top-10 mid-quality predictions.

Using the same potential, we predicted putative binding poses of the bound Test Set and the unbound Test Set (Fig. 7: unbound are shown with lighter color). For both cases, we only show results for the 88 complexes of separation distance less or equal to 30 Å from a total of 230. As can be expected, the predicting accuracy of our docking method is worse for the unbound set. For top 10, mid-quality drops from 44.4 to 20.5% and low quality from 72.2 to 34.1%. Figure 8 shows an example of three correctly predicted complexes from the unbound set with I-RMS < 4 Å. These complexes are classified as difficult in the Protein Docking Benchmark. Nevertheless, PEPSI-Dock produced the correct predictions ranked first, third and sixth. It is interesting to compare the obtained results with the ones produced with other methods. Table 1 presents the comparison of our method on the unbound Test Set with ZDOCK 3.0 (https://zlab.umassmed.edu/zdock/performance.shtml, 2012) and with SwarmDock (Torchala *et al.*, 2013). We should note that ZDOCK uses FFT on a 1.2 Å uniform Cartesian grid and 6° sampling for the rotational DOFs. As for us, we sample the spherical grid of 7.5° angular spacing with 1 Å translational step, which is coarser compared to the one from ZDOCK, especially far away from the origin. ZDOCK defines the success rate by counting the number of complexes found in top 10 predictions with I-RMS $\leq$ 2.5 Å. Table 1 shows that using the same definition of the success rate, PEPSI-Dock slightly outperforms ZDOCK and SwarmDock.

As can be seen from Figure 6, PEPSI-Dock produces significantly better results on the bound complexes than Hex for the same angular sampling step size. However, both methods share the same limitations of a loss of precision due to spherical sampling and a degradation of

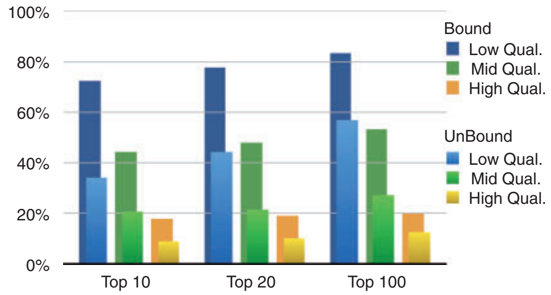

Fig. 7. Percentage of success rates of PEPSI-Dock using I-RMS criteria on the 88 complexes from the Test Set for which the separation distance $\leq$ 30 Å. With darker colours, results are starting from the bound conformations, lighter colours show the unbound results

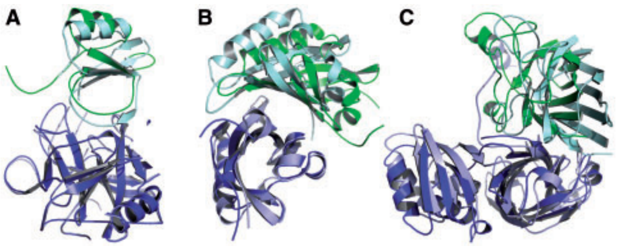

Fig. 8. Examples of the docking predictions for three proteins from the Protein Docking benchmark, classified as difficult. Reference structures of the complexes in the bound form are shown in a light colour. PEPSI-Dock predictions are shown in a dark colour. Left: Pdb code 1ACB, top-1 solution has I-RMS of 2.96 Å. Middle: Pdb code 3F1P, top-3 solution has I-RMS of 3.72 Å. Right: Pdb code 1JK9, top-6 solution has I-RMS of 3.47 Å. For the representation, receptors of docking predictions, coloured in dark blue, were structurally superposed with the receptors of the bound structures coloured in light blue

the polynomial representation at distances larger than 30 Å. One way to remedy such problems would be to cut large proteins into smaller fragments and then to cross-dock the fragments, but this is beyond the scope of the current work.

Overall, looking at the top-10 success rates on the bound complexes, we can observe very good performance of our method on both Training and Test Sets. Even though we did not train the potential to predict near-native or unbound conformations, we still achieve fair results on the unbound Test Set, retrieving 20.5% of the 88 complexes with separation distance less than 30 Å in the top 10 mid-quality solutions. In the experiments on complexes not used for the training, the performance is high for bound cases and much lower for unbound complexes showing that there is a dependence on the training set but no over-fitting. Exploring several definitions of native conformations and the possible inclusion of near-native conformations could improve the predictions on unbound complexes.

Our method compares favourably with Hex, SwarmDock and ZDOCK, even though ZDOCK uses a finer grid search. It is worth noting that PEPSI-Dock predicted three difficult complexes in the unbound Test Set, i.e. those with large conformational changes from the bound state, with I-RMS < 4 Å (Fig. 8). This indicates that PEPSI-Dock provides a useful approach for complexes having large conformational changes.

Regarding the running time of our method, we measured the time of pre-processing computations (step 5 in Fig. 1), the time of FFT-accelerated docking (step 6 in Fig. 1), as well, as the total running time of PEPSI-Dock for all the structures from the version 5 of protein docking benchmark. Figure 9 shows these three timings as a function of the total number of atoms in a complex. We can see that for a half of the tested complexes the total running time is smaller than 12 min. We can also see that the pre-processing time scales linearly with the size of the protein complex and becomes dominating at about 8000 of atoms, whereas the docking time is approximately constant and equals to 7 min. We should add, however, that the pre-processing step in our calculations was not parallelized in the current version of the code, and is subject for further improvement.

Table 1. Number of correct predictions for PEPSI-Dock, ZDOCK 3.0 (with 6° angular sampling) and SwarmDock (full blind mode) tested on the unbound Test Set, depending on the category of the complex

Category (Nb. of Complexes)	ZDOCK.	SwarmDock	PEPSI-Dock
Easy (45)	12	11	13
Medium (15)	1	1	1
Difficult (15)	0	0	0

A prediction is said to be correct if a conformation with I-RMS ≤ 2.5 Å is found in top 10 solutions. Comparison is made only for the complexes for which the separation distance ≤ 30 Å and for the version 4 of the benchmark.

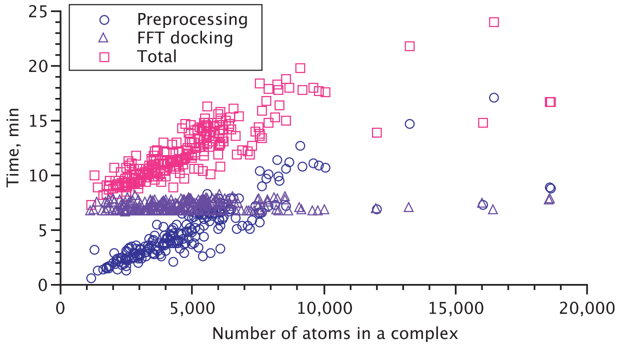

Fig. 9. Running time of PEPSI-Dock as measured on 2 × Intel(R) Xeon(R) CPU E5640 @ 2.67GHz with eight physical cores in total. The pre-processing step corresponds to step 5 in Figure 1. The FFT docking step corresponds to step 6 in Figure 1. The total time corresponds to steps 5–7 in Figure 1

4 Conclusion

We have presented PEPSI-Dock, a new computational method that combines a distant-dependent knowledge-based potential with FFT-accelerated exhaustive sampling on spherical grids. Our potential approximates the binding free energy of protein complexes. We deduce its polynomial expansion coefficients using a training set of protein–protein interfaces and a novel convex optimization problem inspired by a robust machine learning technique. Then, we insert the obtained expansion coefficients into the Hex exhaustive sampling library. This is the first attempt to combine data-driven arbitrary-shaped potentials with a FFT-exhaustive search.

The success rate of our method is particularly high on the bound sets of proteins. However, a limitation of the method is the loss of precision of the spherical sampling grid and the degradation of the polynomial expansions at large distances. On the other hand, results for unbound complexes from the Protein Docking Benchmark are similar to those obtained with Hex, SwarmDock and ZDOCK, despite the fact that we use a coarser sampling grid and our potentials are not specifically trained to predict unbound conformations.

Overall, we have shown that our knowledge-based potential, previously used for the rescoring of protein–protein docking predictions (Popov and Grudinin, 2015) and of protein–ligand docking predictions (Grudinin *et al.*, 2015), can be adapted to exhaustively search the rigid body docking space. Our next developments will be to improve the prediction of unbound structures. To achieve this, we will need to study the influence of the training set on the docking results, and how to choose the most appropriate native and non-native structures. It will also be interesting to construct a hybrid method that combines our knowledge-based scoring function with other sampling methods, particularly with the Cartesian ones, and to develop a multi-centre sampling technique to avoid the loss of precision on the spherical sampling grid when docking very large proteins. Finally, we believe it will be straight-forward to extend our approach to other types of interactions, including protein–ligand and protein–RNA docking problems.

Funding

This work was supported by the Agence Nationale de la Recherche (ANR-11-MONU-006).

Conflict of Interest: none declared.

References

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Böhm,H.J. (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des*., **8**, 243–256.

Bonvin,A.M.J.J. (2006) Flexible protein–protein docking. *Curr. Opin. Struct. Biol*., **16**, 194–200. (

Boyd,S. and Vandenberghe,L. (2004). *Convex Optimization*. Cambridge University Press, New York.

Brooks,B.R. *et al*. (1983) Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem*., **4**, 187–217.

Case,D.A. *et al*. (2005) The amber biomolecular simulation programs. *J. Comput. Chem*., **26**, 1668–1688.

Chae,M.H. *et al*. (2010) Predicting protein complex geometries with a neural network. *Proteins Struct. Funct. Bioinf*., **78**, 1026–1039.

Chaskar,P. *et al*. (2014) Toward on-the-fly quantum mechanical/molecular mechanical (qm/mm) docking: Development and benchmark of a scoring function. *J. Chem. Inf. Model*., **54**, 3137–3152. PMID: 25296988.

Chuang,G.Y. *et al*. (2008) Dars (decoys as the reference state) potentials for protein–protein docking. *Biophys. J*., **95**, 4217–4227.

Eldridge,M.D. *et al*. (1997) Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des*., **11**, 425–445.

Friesner,R.A. *et al*. (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem*., **47**, 1739–1749.

Gabel,J. *et al*. (2014) Beware of machine learning-based scoring functions—on the danger of developing black boxes. *J. Chem. Inf. Model*., **54**, 2807–2815. PMID: 25207678.

Gohlke,H. *et al*. (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol*., **295**, 337–356.

Gray,J. *et al*. (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol*., **331**, 281–300.

Grudinin,S. *et al*. (2015) Predicting binding poses and affinities in the CSAR 2013-2014 docking exercises using the knowledge-based Convex-PL potential. *J. Chem. Inf. Model*., doi:10.1021/acs.jcim.5b00339.

Hess,B. *et al*. (2008) Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput*., **4**, 435–447.

https://zlab.umassmed.edu/zdock/performance.shtml (2012) Performance of ZDOCK on the PDB benchmark 4.0.

Huang,S.Y. and Zou,X. (2008) An iterative knowledge-based scoring function for protein–protein recognition. *Proteins Struct. Funct. Bioinf*., **72**, 557–579.

Huang,S.Y. and Zou,X. (2010) Inclusion of solvation and entropy in the knowledge–based scoring function for protein–ligand interactions. *J. Chem. Inf. Model*., **50**, 262–273. PMID: 20088605.

Huang,S.Y. and Zou,X. (2011) Scoring and lessons learned with the CSAR benchmark using an improved iterative knowledge-based scoring function. *J. Chem. Inf. Model*., **51**, 2097–2106.

Janin,J. (2005) Assessing predictions of protein–protein interaction: the CAPRI experiment. *Protein Sci*., **14**, 278–283.

Katchalski-Katzir,E. *et al*. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U. S. A*., **89**, 2195–2199.

Kinnings,S.L. *et al*. (2011) A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model*., **51**, 408–419. PMID: 21291174.

Koppensteiner,W. and Sippl,M. (1998) Knowledge-based potentials – back to the roots. *Biochemistry*, **63**, 247–252.

Korb,O. *et al*. (2009) Empirical scoring functions for advanced protein–ligand docking with plants. *J. Chem. Inf. Model*., **49**, 84–96.

Kozakov,D. *et al*. (2006) Piper: an FFT-based protein docking program with pairwise potentials. *Proteins Struct. Funct. Bioinf*., **65**, 392–406.

Kuhn,B. *et al*. (2005) Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem*., **48**, 4040–4048. PMID: 15943477.

Lensink,M.F. *et al*. (2016) Prediction of homo- and hetero-protein complexes by ab-initio and template-based docking: a CASP-CAPRI experiment. *Proteins Struct. Funct. Bioinf*. doi: 10.1002/prot.25007.

Li,G.B. *et al*. (2013) ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein–ligand interactions. *J. Chem. Inf. Model*., **53**, 592–600. PMID: 23394072.

Liu,J. and Wang,R. (2015) Classification of current scoring functions. *J. Chem. Inf. Model*., **55**, 475–482. PMID: 25647463.

Maiorov,V.N. and Grippen,G.M. (1992) Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol*., **227**, 876–888.

Méndez,R. *et al*. (2003) Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins Struct. Funct. Genet*., **52**, 51–67.

Mooij,W. and Verdonk,M.L. (2005) General and targeted statistical potentials for protein–ligand interactions. *Proteins Struct. Funct. Bioinf*., **61**, 272–287.

Muegge,I. and Martin,Y.C. (1999) A general and fast scoring function for protein–ligand interactions: a simplified potential approach. *J. Med. Chem*., **42**, 791–804.

Nocedal,J. and Wright,S. (2000) *Numerical Optimization. Springer Series in Operations Research and Financial Engineering*. Springer, New York.

Popov,P. and Grudinin,S. (2014) Rapid determination of RMSDs corresponding to macromolecular rigid body motions. *J. Comput. Chem*., **35**, 950–956.

Popov,P. and Grudinin,S. (2015) Knowledge of native protein–protein interfaces is sufficient to construct predictive models for the selection of binding candidates. *J. Chem. Inf. Model*, **55**, 2242–2255.

Qiu,J. and Elber,R. (2005) Atomically detailed potentials to recognize native and approximate protein structures. *Proteins Struct. Funct. Bioinf*., **61**, 44–55.

Ritchie,D. (2005) High-order analytic translation matrix elements for real-space six-dimensional polar Fourier correlations. *J. Appl. Crystallogr*., **38**, 808–818.

Ritchie,D.W. and Kemp,G. (2000) Protein docking using spherical polar Fourier correlations. *Proteins Struct. Funct. Genet*., **39**, 178–194.

Ritchie,D.W. and Venkatraman,V. (2010) Ultra-fast FFT protein docking on graphics processors. *Bioinformatics*, **26**, 2398–2405.

Ritchie,D.W. *et al*. (2008) Accelerating and focusing protein–protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics*, **24**, 1865–1873.

Scholkopf,B. and Smola,A.J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Tobi,D. and Bahar,I. (2006) Optimal design of protein docking potentials: efficiency and limitations. *Proteins Struct. Funct. Bioinf*., **62**, 970–981.

Torchala,M. *et al*. (2013) Swarmdock: a server for flexible protein–protein docking. *Bioinformatics*, **29**, 807–809.

Vreven,T. *et al*. (2015) Updates to the integrated protein–protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*., **427**, 3031–3041.

Wang,R. *et al*. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des*., **16**, 11–26.

Zhou,H. and Skolnick,J. (2011) Goap: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J*., **101**, 2043–2052.

Zilian,D. and Sotriffer,C.A. (2013) SFCscoreRF: a random forest-based scoring function for improved affinity prediction of protein–ligand complexes. *J. Chem. Inf. Model*., **53**, 1923–1933. PMID: 23705795.