TASK QUARTERLY 18 No 3, 219–229

# COARSE-GRAINED MODELING OF PROTEIN STRUCTURE, DYNAMICS AND PROTEIN-PROTEIN INTERACTIONS

ANDRZEJ KOLINSKI, SEBASTIAN KMIECIK, MICHAL JAMROZ, MACIEJ BŁASZCZYK, MAKSIM KOUZA AND MATEUSZ KURCINSKI

*Laboratory of Theory of Biopolymers, Faculty of Chemistry*
*University of Warsaw*
*Pasteura 1, 02-093 Warsaw, Poland*

(Paper presented at the CBSB14 Conference, May 25–27, 2014, Gdansk, Poland)

**Abstract:** Theoretical prediction of protein structures and dynamics is essential for understanding the molecular basis of drug action, metabolic and signaling pathways in living cells, designing new technologies in the life science and material sciences. We developed and validated a novel multiscale methodology for the study of protein folding processes including flexible docking of proteins and peptides. The new modeling technique starts from coarse-grained large-scale simulations, followed by selection of the most plausible final structures and intermediates and, finally, by an all-atom rectification of the obtained structures. Except for the most basic bioinformatics tools, the entire computational methodology is based on the models and algorithms developed in our lab. The coarse-grained simulations are based on a high-resolution lattice representation of protein structures, a knowledge based statistical force field and efficient Monte Carlo dynamics schemes, including Replica Exchange algorithms. This paper focuses on the description of the coarse-grained CABS model and its selected applications.

**Keywords:** coarse-grained modeling, protein folding, protein dynamics, molecular docking, protein docking

## 1. Background

Numerous genomic projects provide a plethora of protein sequences. The number of experimentally solved protein structures is about a thousand times (or at least several hundred times) smaller. The reason for this still increasing gap is simple. The sequencing of genomic materials is usually highly automated and relatively inexpensive. On the contrary, the determination of protein structures by means of X-ray crystallography, NMR and, to a lesser extent, by other experimental techniques, is very expensive, time consuming, and requires high-qualified personnel. The knowledge of protein structures is essential for protein

function prediction, rational drug design, elucidation of complex evolutionary relationships and modeling complex networks of dynamics and interactions in a living cell [1, 2] at the molecular level, including conformational transitions and the aggregation of proteins linked to various (especially neurodegenerative) diseases. Protein structure determination is also necessary for protein engineering in biotechnology and modern material science. *In silico* modeling support of experimental structure determination is equally important. It is possible to use the sparse experimental data (from: NMR, fluorescence, cross-linking, Electron Microscopy, etc.) as restraints for advanced molecular modeling [3]. A good *in silico* model can be used in molecular replacements in X-ray crystallography, facilitating crystallographic structure determination. For a fraction of known sequences depending on the level of sophistication of the methods used, *in silico* prediction of protein structures is now feasible, and sometimes provides structures of a similar quality to those obtained experimentally [4]. It is extremely important to be able to model as large a fraction of all proteins as possible [5, 6] with high accuracy. Thus, any progress in high fidelity protein structure prediction would have a large impact on various areas of life sciences and biotechnology. Experimental determination of protein interactions (and assemblies of proteins with other molecules) is even less advanced [2, 7], as theoretical predictions of protein assemblies are more difficult due to a significantly higher complexity of the multi-macromolecular systems. Dependable, routine prediction of protein associates (and possibly elements of molecular pathways leading to the active structures) is a challenging next step (after the structure prediction of single molecules) in the deciphering of the molecular mechanisms of life [2, 7].

Recently, it has been demonstrated that not only structure prediction, but also the prediction of protein folding mechanisms and the mechanisms of biomacromolecular association could be effectively studied using coarse-grained models, relying either on a physics-based interaction scheme (UNRES model [8–10]) or a knowledge-based force-field (CABS model [11–13]). The reduction of the number of explicitly treated degrees of freedom and the smoothening of protein free-energy landscapes (by a properly designed potential of the mean-force of statistical origin) lead to a speed-up of a protein fold assembly by orders of magnitude [11]. This is essential, since the real proteins fold in time frames of microseconds to minutes [14]. Presently, the state-of-art all-atom simulations of proteins can be performed for a time frame range of nanoseconds [15]. Thus, properly designed multiscale simulations are now the only way for the large-scale biomacromolecular modeling [12, 13].

In the last few years, we have developed a coarse-grained protein modeling tool: the CABS model [16]. The main objective behind its design was to satisfy the two seemingly contradictory assumptions: a detailed description of the system and computational efficiency. As a result, we came up with a model, which is approximately three to four orders of magnitude faster than the classical Molecular Dynamics, and nonetheless capable of reproducing structural models of proteins

with a resolution comparable to the experimentally determined crystallographic structures. The CABS computational technology has been tested during several editions of CASP (Critical Assessment of Techniques for Protein Structure Prediction, http://predictioncenter.org) world-wide experiment [4]. The blind prediction made by the Kolinski-Bujnicki group during CASP6 ranked as the second one among over 200 world-leading groups participating, and the group was the best when the consistency of the prediction was used as a criterion (the fraction of the constructed models that were placed among the top 20 best predictions). The CABS technology proved to be equally efficient when applied to comparative modeling, difficult fold recognition tasks, as well as in de novo prediction of new folds (lacking structural analogs in the database of already solved structures). Interestingly, the CABS simulations are surprisingly accurate when it comes to reproducing protein folding thermodynamics, dynamics of denatured proteins and protein folding pathways [11, 12] (see Figure 1) and flexible molecular docking of proteins and peptides [13, 17, 18], despite the fact that the force field is based on structures of folded proteins. Perhaps, molecular interactions in denatured proteins are more similar to the interactions seen in folded structures than it was believed in the past.

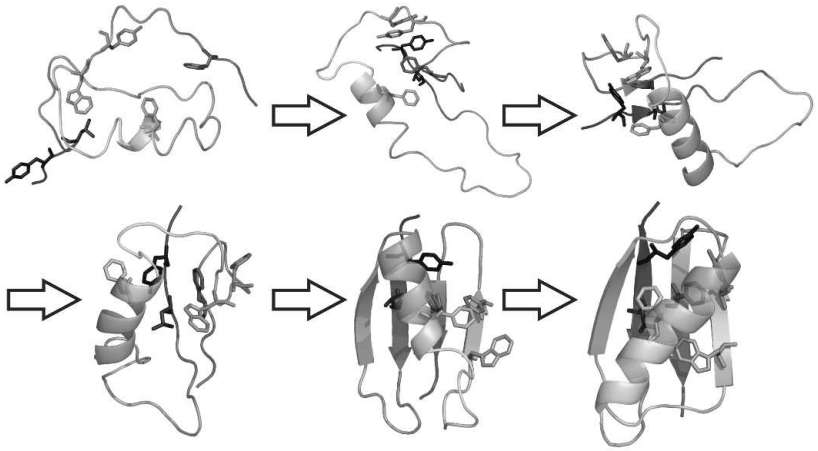

**Figure 1.** Selected snapshots from the *in silico* folding of a small single domain protein (B1 domain of IgG binding protein G) using the CABS algorithm (followed by rebuilding and all-atom minimization procedures), presenting key side-chains interactions (folding nucleus) at various folding stages [12]; heavy atom bonds of the nucleus residues are marked with sticks; the secondary structure is depicted using transparent ribbons; the characteristic features of the molten globule state are visible in the fourth and fifth snapshots; the sequence is the only protein-specific input for the simulation program

## 2. CABS model

The design of the CABS representation, the force field, and the sampling schemes is a result of many years of experience gained by the development of

several reduced-space models of various resolutions [16, 19–22]. The acronym CABS stands for united atoms representing a residue in a polypeptide chain: $\alpha$-carbon (CA), $\beta$-carbon (B) and center of mass of the side group (S), where applicable. Additionally, one more pseudo-atom is defined in the peptide bond center to support a model of main-chain hydrogen bonds, essential for the regularization of the geometry of a protein structure. The force field of the CABS has a knowledge-based character and the potentials of mean-force, mimicking the physical interactions in proteins, were derived from a careful analysis of structural regularities present in the known high-resolution protein structures. For instance, the square-well type potential, describing interactions between the side chains, reflects the frequency of contacts between pairs of residues of various types. The potentials are generated using the standard Boltzmann inversion, and the cut-off distances are also pairwise specific. Furthermore, the side-group potentials are context-dependent. Namely, the strength of the interactions depends on the mutual orientation of the contacting side groups and on the local geometry of the main chain fragments involved. In this way, very complex multibody correlations including averaged solvent effects, are accounted for in an implicit fashion. This feature of the CABS force field distinguishes it from dozens of other statistical potentials described in the literature, and leads to a much higher sequence specificity. Let us give just one illustration of this novel approach. In the CABS force field, two oppositely charged groups (LYS-GLU for instance) attract each other when oriented in an approximately parallel fashion (the vector from $C_\alpha$ to the center of mass of a given side group defines its orientation), while they are repulsive when in the antiparallel orientation. This feature reflects the fact that the polar groups in globular proteins are located almost exclusively on the protein surface, thus, the near-by side chains must be approximately parallel. On the contrary, burring the two charged residues inside the globule would usually lead to antiparallel contact, and is energetically unfavorable (therefore very rarely observed in folded proteins). The pairwise side chain potentials are the main component of the long-range non-bonded interactions of the model. The short range interactions are also of the statistical origin and contain two types of components: generic protein-like biases, which favor protein-like local geometry of the $C_\alpha$-trace, and sequence dependent potentials controlling possible geometries of the five-residue fragments. The model of hydrogen bonds is based on a translation of the geometry of the main chain hydrogen bonds onto a set of geometric requirements applied to the corresponding fragments of the $C_\alpha$-trace. This ersatz of the main chain hydrogen bonds exhibits a high correlation (above 95%) with the all-atom DSSP definition.

The CABS employs high-resolution lattice discretization of the protein conformational space (see Figure 2). The positions of $\alpha$-carbons ($C_\alpha$) are restricted to a simple cubic lattice, with the grid spacing equal to 0.61 Å. Thus the average accuracy of the projection of the crystallographic $C_\alpha$ coordinates onto the lattice has the range of 0.35 Å. $C_\alpha$-$C_\alpha$ virtual bonds, slightly fluctuating around the

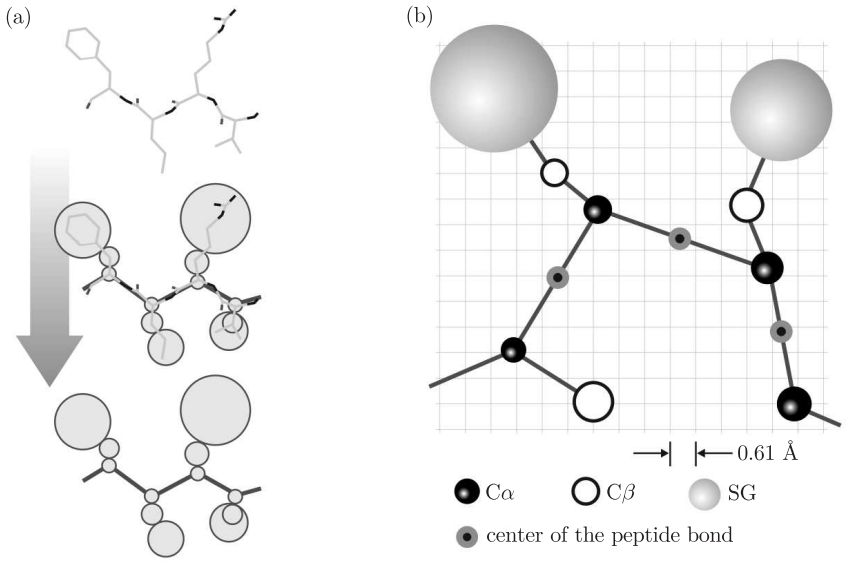


**Figure 2.** Schematic illustration of protein representation in the CABS model: (a) reduction of the degrees of freedom; (b) lattice representation of a fragment of a peptide chain; the sizes of the spheres representing united atoms do not correspond to the real volumes

equilibrium length, belong to the set of 800 lattice vectors. Thus, the lattice artifacts could be safely ignored. The positions of the remaining united atoms ($\beta$-carbons, side-chains and pseudo-atoms representing the peptide bond units) are defined in the local coordinate systems defined by the $C_\alpha$-trace, and are not restricted to the lattice. Two rotamers only are possible for each side chain. Although being an obvious limitation, such solution is at the same time very beneficial for the speed of the computations. It needs also to be pointed out that the reduced representation of CABS is consistent with the all atom representation. The lattice models could be quite accurately reconstructed into the complete structures [6, 23].

Various Monte Carlo schemes are used for efficient sampling of the CABS conformational space. When the main task is to find the native-like structure, different variants of the Replica Exchange Monte Carlo are used. Single copy Metropolis schemes (isothermal or within a simple simulated annealing scheme) are employed for the studies of protein dynamics and folding mechanisms. A properly designed set of local random micromodifications leads to the numerical solution of the Master Equation of motion, and thereby provides a reasonable picture of the system dynamics (except for the short time-scale comparable to the average time of the local conformational transitions).

It should be pointed out that the lattice representation permits extremely fast conformational updating – random mechanism simply references to large tables of possible local conformations. Moreover, the computation of the system's conformational energy is highly simplified by the lattice representation. As a result, the CABS sampling is about two orders of magnitude faster than it would

be for the otherwise equivalent continuous space model. The speed-up in respect to all-atom molecular dynamics is a range of five orders of magnitude. Therefore, it is possible to fold small proteins, starting from a random conformation, in span of minutes of a single LINUX box CPU time. Designing „smart" collective local micromodifications (planned for the next update of CABS) should permit an additional 10-fold speed-up of the structure assembly.

## 3. Multiscale modeling

For the purpose of drug design and computer-aided protein engineering, as well as the design of new artificial proteins for biotechnology, it is necessary to build models with all-atomic details. The all-atom models are also very useful in the selection of the best models from the CABS simulations [6]. The all-atom force fields are better correlated with the distance from the native structure in the close vicinity of the native state, while the CABS force field yields better results for highly distorted decoys. The CABS representation has a sufficient resolution for a dependable and quite accurate reconstruction of all-atom structures. We developed our own suite of software for a very fast in-flow transitions between the reduced and all-atom representation [23]. The idea is outlined in Figure 3.

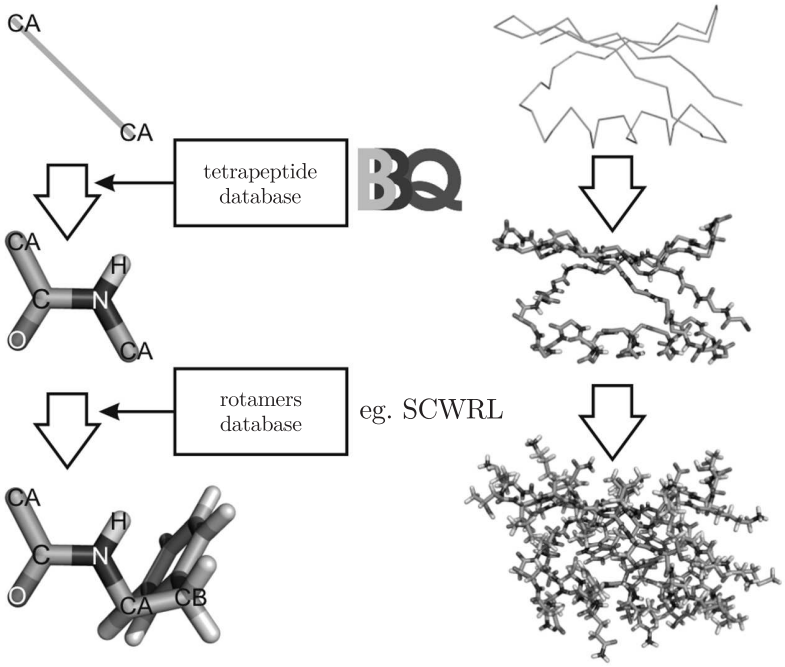


**Figure 3.** Illustration of the two-stage rebuilding of an all-atom structure from a coarse-grained CABS representation; first, the coordinates of the main chain (and the $\beta$-carbon atoms) are reconstructed by an extremely fast and accurate BBQ algorithm [23] developed in our lab; then, the side groups are placed using, for example, the SCWRL algorithm [24]

The test reduction (projection onto the CABS representation) of the high-resolution crystallographic structures followed by the all-atom reconstruction produces only 0.3 Å error on the main chain and about 1.0 Å error on the full-detail structures [16]. Interestingly, the all atom reconstruction could be dependably applied to the partially folded proteins and folding intermediates [12]. Thus, going back and forth between the reduced and all-atom simulations facilitates the studies in the atomic resolution of the large-scale (in respect to the system size and the range of structure relaxation) processes in biomacromolecular systems.

## 4. Molecular protein-protein docking simulations

A multichain version of the CABS model enables the simulation of large assemblies of proteins and peptides. Several algorithms based on CABS have been tested and the results are very encouraging [17, 18, 25, 26] (see Figures 4–5). What is important, the docking of peptides to proteins could be done without any knowledge of the final structure of the peptide and the location of the docking interface. Significant flexibility of the entire protein structure could be allowed during the docking simulations. This already goes far beyond the present state-of-art molecular docking computational technology [27–30].

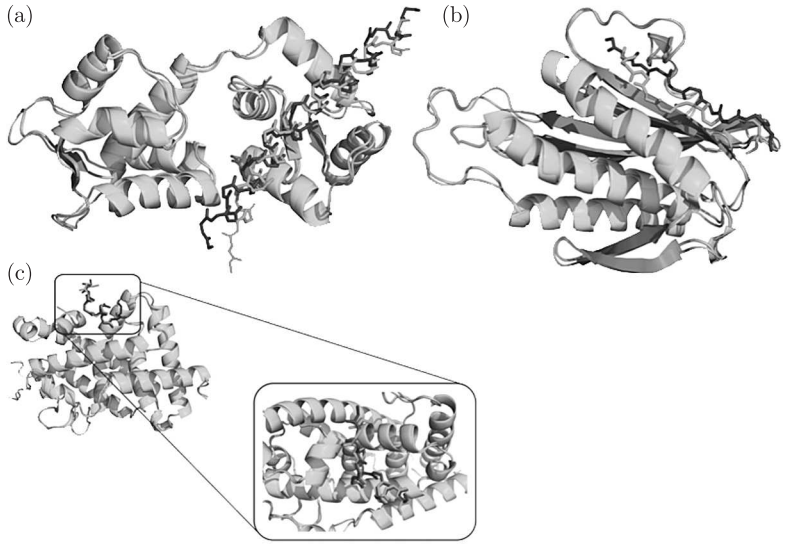


**Figure 4.** The docking results of fully flexible peptides to semi-flexible receptors; nothing was assumed about the final conformation and the location of the peptides in respect to the receptor proteins; case (c) illustrates the docking of a very important activator to vitamin D receptor [18]; the inset shows the details of the assembled structure (shown from a different direction) focused on the activator; proteins are shown in cartoons, peptides in sticks. Peptide models (light gray) superimposed onto crystallographic structure (dark gray); PDB codes: 2A2X (a), 1KLQ (b), 1RJK (c)

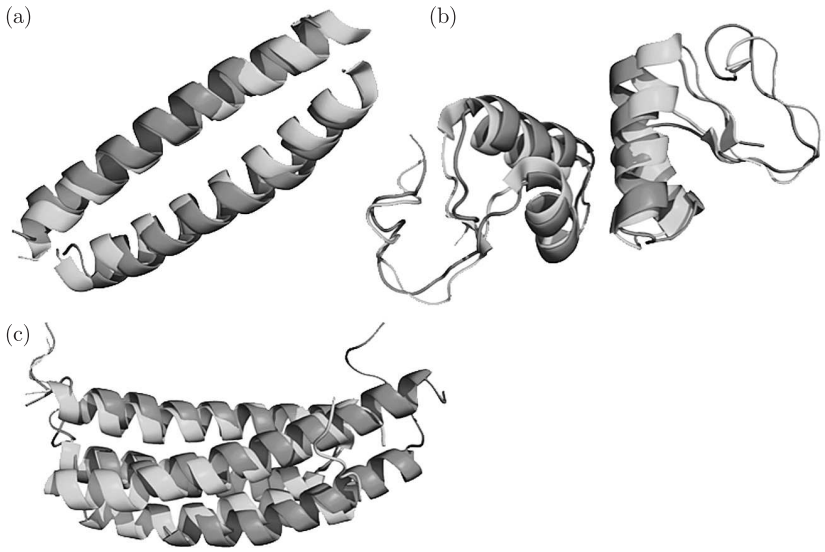


**Figure 5.** Docking results for small protein homodimers [17]; (a) a fully flexible assembly of the GCN4 leucine zipper; (b)–(c) docking of a fully flexible (unrestricted conformational space) protein to a semi-flexible protein (weak native-like restraints superimposed); crambin-like dimer for (b) and ROP dimer for (c), respectively; models (light gray) superimposed onto a crystallographic structure (dark gray); PDB codes: 2ZTA (a), 1OKH (b), 1RPR (c)

## 5. Modeling of folding and binding of intrinsically disordered protein

Phosphorylated kinase-inducible domain (pKID) is a small protein which lacks a well-defined three-dimensional structure in the isolated state. Although, when binding with its interacting domain (KIX), the pKID adopts a specific three-dimensional structure. The mechanism of such induced binding and folding process is not clearly understood. We studied this process by employing free docking simulations. The KIX structure was treated as partially flexible, oscillating during simulation near its native structure. The pKID was treated as a completely free object, without any assumptions about its tertiary structure, and any information about the binding site. Several replica exchange Monte Carlo dynamics simulations with CABS model were performed. Figure 6 shows an example starting structure for such simulations. The receptor structure starts from a near native state, while several random conformations of the pKID chains are placed in the vicinity. During the simulations the copies of the pKID are not visible to each other, and are present only for speeding up the search procedures for two molecules of the pKID-KIX system. The obtained transient encounter complexes on the path to native binding are illustrated in Figure 7 and in Figure 8. A detailed analysis of the simulation results clearly indicates a nucleation-condensation mechanism of a pKID structure assembly, which is very similar for a common scenario of a globular proteins folding process.

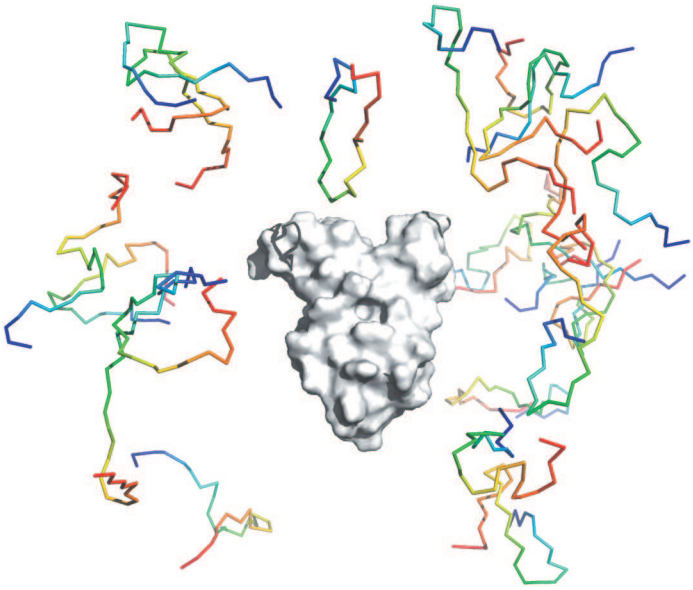

**Figure 6.** Folding and binding mechanism of a disordered pKID peptide; alternative starting conformations and positions of the pKID peptide are shown in color, while the KIX domain is shown as a gray surface

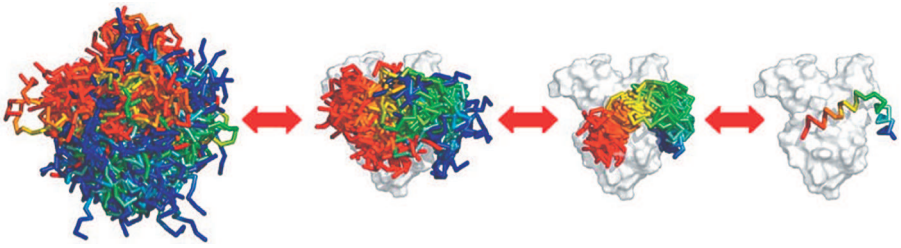

**Figure 7.** Folding and binding mechanism of a disordered pKID peptide; alternative pKID peptide conformations observed during simulations are presented in color, while the KIX domain is shown as a gray surface

## 6. Conclusions

Coarse-grained modeling, especially when combined with all-atom refinement and ranking of the obtained models is a very powerful method for protein structure prediction, study of protein dynamics and flexible, unrestrained molecular docking of peptides and proteins. The modeling tools developed in our lab are easily available for scientific community. This includes: CABS-fold (http://biocomp.chem.uw.edu.pl/CABSfold/) – a web server for de novo prediction of protein structures [31], CABS-flex (http://biocomp.chem.uw.edu.pl/CABSflex/) – a server for simulations of near-native dynamics of proteins [32, 33] and CABS-dock (http://biocomp.chem.uw.edu.pl/tools/cabsdock) – a server for flexible protein-peptide docking (article in preparation). Additional useful tools, such as the CABS-oriented python libraries PyCABS [34] and Bioshell [35, 36] and

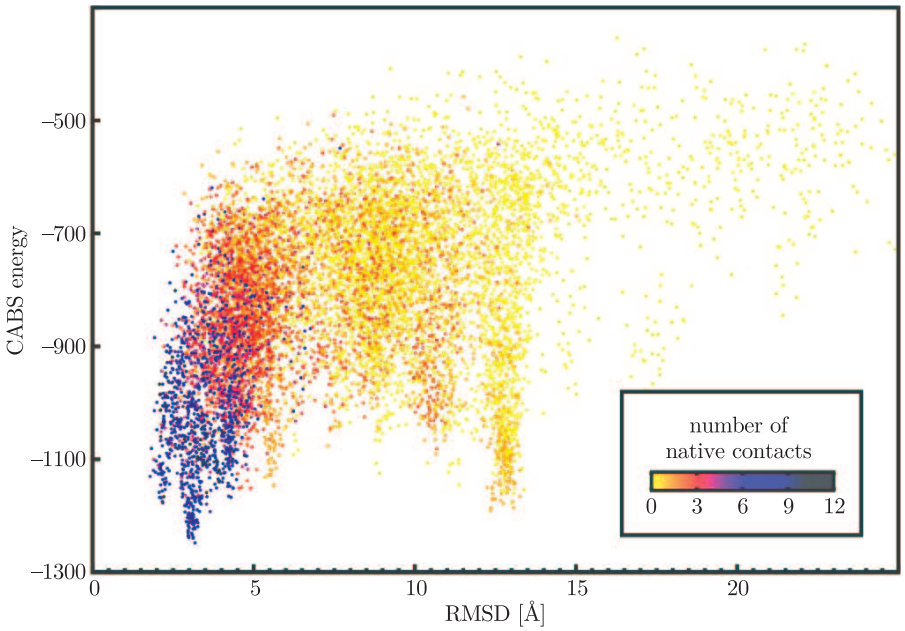


**Figure 8.** Folding and binding mechanism of a disordered pKID peptide; the plot shows CABS energy vs. resemblance to the native complex (RMSD) for protein models from the example folding and binding trajectory

Clusco – program for comparison and clustering of protein structures [37], are also available for download at http://biocomp.chem.uw.edu.pl/tools.

## *Acknowledgements*

We acknowledge the funding from the Foundation for the Polish Science TEAM project [TEAM/2011–7/6] cofinanced by the EU European Regional Development Fund operated within the Innovative Economy Operational Program and from the Polish National Science Centre (NCN), Grant No. DEC-2011/01/D/NZ2/07683, and from Polish Ministry of Science and Higher Education, Grant No. IP2012 016872.

## *References*

[1] Laskowski R A and Thornton J M 2008 *Nat. Rev. Genet.* **9** (2) 141
[2] Wolfson H J *et al.* 2005 *Curr. Protein Pept. Sci.* **6** (2) 171
[3] Latek D, Ekonomiuk D and Kolinski A 2007 *J. Comput. Chem.* **28** (10) 1668
[4] Kolinski A and Bujnicki J M 2005 *Proteins* **61** 84
[5] Bradley P, Misura K M and Baker D 2005 *Science* **309** (5742) 1868
[6] Kmiecik S, Gront D and Kolinski A 2007 *BMC Struct. Biol.* **7** 43
[7] Schueler-Furman O. *et al.* 2005 *Science* **310** (5748) 638
[8] Liwo A, Khalili M and Scheraga H A 2005 *Proc. Natl. Acad. Sci. USA* **102** (7) 2362
[9] Oldziej S *et al.* 2005 *Proc. Natl. Acad. Sci. USA* **102** (21) 7547
[10] Sieradzan A K, Liwo A and Hansmann U H 2012 *Journal of Chemical Theory and Computation* **8** (9) 3416
[11] Kmiecik S and Kolinski A 2007 *Proc. Natl. Acad. Sci. USA* **104** (30) 12330

[12] Kmiecik S and Kolinski A 2008 *Biophys. J.* **94** (3) 726
[13] Kurcinski M, Kolinski A and Kmiecik S 2014 *J. Chem. Theory Comput.* **10** (6) 2224
[14] Lindorff-Larsen K. *et al.* 2011 *Science* **334** (6055) 517
[15] Klepeis J L *et al.* 2009 *Curr. Opin. Struct. Biol.* **19** (2) 120
[16] Kolinski A 2004 *Acta Biochim Pol.* **51** (2) 349
[17] Kurcinski M and Kolinski A 2007 *J. Mol. Model* **13** ((6–7)) 691
[18] Kurcinski M and Kolinski A 2007 *J. Steroid Biochem. Mol. Biol.* **103** ((3–5)) 357
[19] Kolinski A, Skolnick J and Yaris R 1986 *Proc. Natl. Acad. Sci. USA* **83** (19) 7267
[20] Skolnick J and Kolinski A 1990 *Science* **250** (4984) 1121
[21] Skolnick J, Kolinski A and Ortiz A R 1997 *J. Mol. Biol.* **265** (2) 217
[22] Kolinski A *et al.* 2001 *Proteins* **44** (2) 133
[23] Gront D, Kmiecik S and Kolinski A 2007 *J. Comput. Chem.* **28** (9) 1593
[24] Canutescu A A, Shelenkov A A and Dunbrack R L Jr. 2003 *Protein Sci.* **12** (9) 2001
[25] Horwacik I *et al.* 2011 *Int. J. Mol. Med.* **28** (1) 47
[26] Steczkiewicz K *et al.* 2011 *Proc. Natl. Acad. Sci. USA* **108** (23) 9443
[27] Ritchie D W 2008 *Curr. Protein Pept. Sci.* **9** (1) 1
[28] Bonvin A M 2006 *Curr. Opin. Struct. Biol.* **16** (2) 194
[29] Wang C, Bradley P and Baker D 2007 *J. Mol. Biol.* **373** (2) 503
[30] Lensink M F and Mendez R 2008 *Curr. Pharm. Biotechnol.* **9** (2) 77
[31] Blaszczyk M, *et al.* 2013 *Nucleic Acids Res.* **41**, W406 (Web Server issue)
[32] Jamroz M, Kolinski A and Kmiecik S 2014 *Bioinformatics* **30** (15) 2150
[33] Jamroz M, Kolinski A and Kmiecik S 2013 *Nucleic Acids Res.* **41**, W427 (Web Server issue)
[34] Jamroz M, Kolinski A and Kmiecik S 2014 *Methods Mol. Biol.* **1137** 235
[35] Gront D and Kolinski A 2008 *Bioinformatics* **24** (4) 584
[36] Gront D and Kolinski A 2006 *Bioinformatics* **22** (5) 621
[37] Jamroz M and Kolinski A 2013 *Bmc Bioinformatics* **14** 62