UNIWERSYTET WARSZAWSKI
WYDZIAŁ CHEMII

Development of multiscale methods for protein molecular modeling and their application in studies of the mechanisms of protein function

dr Sebastian Kmiecik

Warsaw, October 2014

Summary of professional accomplishments – Attachment No. 1 (English version) to the application for the habilitation qualification

Uniwersytet Warszawski, Wydział Chemii – ul. Pasteura 1, 02-093 Warszawa
tel. 022 822 09 75; centr. 022 822 02 11; fax 022 822 59 96
e-mail: dziekan@chem.uw.edu.pl www: http://www.chem.uw.edu.pl
Bank Millenium S.A. 12 1160 2202 0000 0000 6084 9173 NIP 525-001-12 -66

Summary of professional accomplishments

1. Name and surname

Sebastian Kmiecik

2. Held diplomas, scientific / arts degrees - with the name, place and year of acquisition, and the title of doctoral dissertation.

09.2007 – Doctoral degree in chemistry, specialization: theoretical chemistry. Faculty of Chemistry, University of Warsaw. Title of the doctoral thesis "Multiscale modeling of folding mechanisms of globular proteins", supervisor: prof. dr hab. Andrzej Koliński.

09.2002 – Master of Science degree in chemistry, Faculty of Chemistry, University of Warsaw.

3. Information on current and previous employment in scientific /art institutions.

from 03.2010 till now - Faculty of Chemistry, University of Warsaw.

Position: specialist in science

from 10.2007 to 02.2010 – Selvita, Kraków.

Position: Head of the department of computational chemistry (from 02.2009 to 02.2010); Project manager (from 02.2008 to 01.2009); System architect (from 10.2007 to 01.2008).

Sevita is a research and development company providing comprehensive solutions in chemistry, biotechnology and bioinformatics to research and development units and various branches of industry. When working for Selvita, I performed scientific studies and also implemented innovative solutions in computer-aided pharmaceutical and biotechnological research:

- as a manager of research and implementation projects (projects positively assessed and funded by the Innovative Economy Operational Programme of the European Union and by the Chief Technical Organization) – as described in the Attachment No. 2 of this application; and of research projects in computer-aided drug design.
- as a participant (lecturer or poster presenting person) of international scientific conferences, workshops and trainings – as described in the Attachment No. 2 of this application.

4. Indication of achievement[1] under Art. Paragraph 16. 2 of the Act of 14 March 2003 Academic Degrees and Title, and Degrees and Title in Art. (Dz. U. No 65, item. 595 with amendments):

a) the title of the scientific achievement:

"Development of multiscale methods for protein molecular modeling and their application in studies of the mechanisms of protein function"

b) publications comprising the academic achievement

* – corresponding author

H1. Michał Jamróz, Modesto Orozco, Andrzej Koliński and **Sebastian Kmiecik***. *Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field*. **Journal of Chemical Theory and Computation**, 9: 119-125, 2013.

H2. Michał Jamróz, Andrzej Koliński and **Sebastian Kmiecik***. *CABS-flex: server for fast simulation of protein structure fluctuations.* **Nucleic Acids Research**, 41: W427-W431, 2013.

H3. Michał Jamróz, Andrzej Koliński and **Sebastian Kmiecik***. *CABS-flex predictions of protein flexibility compared with NMR ensembles.* **Bioinformatics**, 30(15):2150-4, 2014.

H4. Maciej Błaszczyk, Michał Jamróz, **Sebastian Kmiecik*** and Andrzej Koliński*. *CABS-fold: server for the de novo and consensus-based prediction of protein structure*. **Nucleic Acids Research**, 41:W406-W411, 2013.

H5. Sebastian Kmiecik, Michał Jamróz and Michał Koliński*. *Structure prediction of the second extracellular loop in G-protein-coupled receptors.* **Biophysical Journal**, 106(11):2408-16, 2014.

[1]if this is the achievement of joint publication / publications, provide a declaration of all its coauthors, determining the individual contribution of each of them in its creation.

H6. **Sebastian Kmiecik*** and Andrzej Koliński*. *Simulation of chaperonin effect on protein folding: a shift from nucleation-condensation to framework mechanism.* **Journal of the American Chemical Society**, 133: 10283–9, 2011.

H7. **Sebastian Kmiecik***, Dominik Gront, Maksim Kouza and Andrzej Koliński*. *From coarse-grained to atomic-level characterization of protein dynamics: transition state for the folding of B domain of protein A*. **The Journal of Physical Chemistry B**, 116: 7026–32, 2012.

H8. Mateusz Kurciński, Andrzej Koliński and **Sebastian Kmiecik***. *Mechanism of Folding and Binding of an Intrinsically Disordered Protein as Revealed by Ab Initio Simulations.* **Journal of Chemical Theory and Computation**, 10 (6): 2224–31, 2014.

H9. Jacek Wabik, **Sebastian Kmiecik**, Dominik Gront, Maksim Kouza* and Andrzej Koliński*. *Combining Coarse-Grained Protein Models with Replica-Exchange All-Atom Molecular Dynamics.* **International Journal of Molecular Sciences**, 14: 9893–9905, 2013.

H10. Michał Jamroz, Andrzej Kolinski and **Sebastian Kmiecik***. *Protocols for efficient simulations of long time protein dynamics using coarse-grained CABS model*. **Methods in Molecular Biology** (Clifton, N.J.), 1137:235-250, 2014.

H11. **Sebastian Kmiecik**, Michał Jamroz and Andrzej Koliński*. *Multiscale approach to protein folding dynamics*. Rozdział w książce: **Multiscale Approaches to Protein Modeling**, Springer Berlin Heidelberg, Andrzej Kolinski, ed., 281-294, 2011.

H12. **Sebastian Kmiecik***, Jacek Wabik, Michał Koliński, Maksim Kouza and Andrzej Koliński. *Coarse-Grained Modeling of Protein Dynamics*. Rozdział w książce: **Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes** – Springer Series in Bio-/Neuroinformatics, Springer Berlin Heidelberg, Adam Liwo, ed., 1:55-79, 2014.

c) discussion of the scientific / artistic goals of the above publication / publications and the results achieved together with a discussion of their possible use

Introduction to the subject and the scientific goals of the publications comprising the academic achievement

Proteins are fundamental molecules involved in basic functions of living organisms. Understanding protein functions on the molecular level requires characterization of their structure and dynamics. In vivo, the level of structural mobility depends on a protein and may lie on a continuum between the two extremes: rigid or almost rigid structure (e.g. structures of most of the enzymes) and a highly mobile disordered one (Chouard 2011).

So far, the molecular basis of protein function has been best explored for proteins that have high structural rigidity. The least is known about the molecular mechanisms of action of proteins that exhibit high structural variability. The degree of difficulty of scientific studies, both experimental and theoretical, increases with the degree of mobility and the size of the studied proteins.

The major goal of the publications comprising the academic achievement was the development of new and efficient methods for protein molecular modeling and their application in studies of the mechanisms of protein function. These publications were related mainly to the development of new coarse-grained methodologies applicable in efficient simulations of protein dynamics, but they also concerned application and optimization of various all-atom modeling strategies in combination with coarse-grained modeling.

Both protein structure and protein dynamics can by characterized hierarchically (Henzler-Wildman, Lei et al. 2007). Therefore, protein modeling can be performed by multiscale modeling i.e. on various levels of protein structure representation and dynamic time scales (see review papers **H11** and **H12**). In the last few decades, there has been an intensive development of protein modeling techniques, both in coarse-grained and all-atom resolutions. All-atom molecular dynamics (MD) is presently considered as a golden standard in the field of protein simulations (Karplus and Kuriyan 2005). Unfortunately, despite the rapid increase in computing power, applications of all-atom MD are still limited to short time scales (to a microsecond time scale when using general-purpose supercomputers, or a millisecond time scale for small proteins when using MD dedicated supercomputer (Shaw, Maragakis et al. 2010; Shaw 2013)). Therefore, most of the conformational changes responsible for important protein functions remain inaccessible to all-atom MD due to their too large time scales and/or a too large size of the modeled system (Vendruscolo and Dobson 2011). Efficient simulations of processes involving large time scales and system sizes may be achieved using coarse-grained models (see also review papers **H11**, **H12** and (Liwo, He et al. 2011)). Recently, there has been a dramatic increase in the number of studies that rely on coarse-grained simulations (Takada 2012). That increase is largely due to the intensification of experimental studies, which frequently deliver only sparse data about the structure and dynamics of the systems studied. Therefore, the major role of molecular modeling methods is interpretation of experimental data by providing all-atom molecular models, which describe the modeled processes. The transition between efficient simulation in coarse grained representation to results in all-atom resolution is possible thanks to the multiscale modeling approach (see Figure 1).

Last year, the important role of multiscale modeling approach in contemporary biochemical research has been acknowledged by the Nobel prize committee. The Committee awarded the 2013 Prize in Chemistry “for the development of multiscale models for complex chemical systems”, recognizing the work of Michael Levitt and Arieh Warshel in protein

coarse-grained modeling (Levitt and Warshel 1975; Warshel and Levitt 1976) as an important step in the investigation of large biomolecular systems.

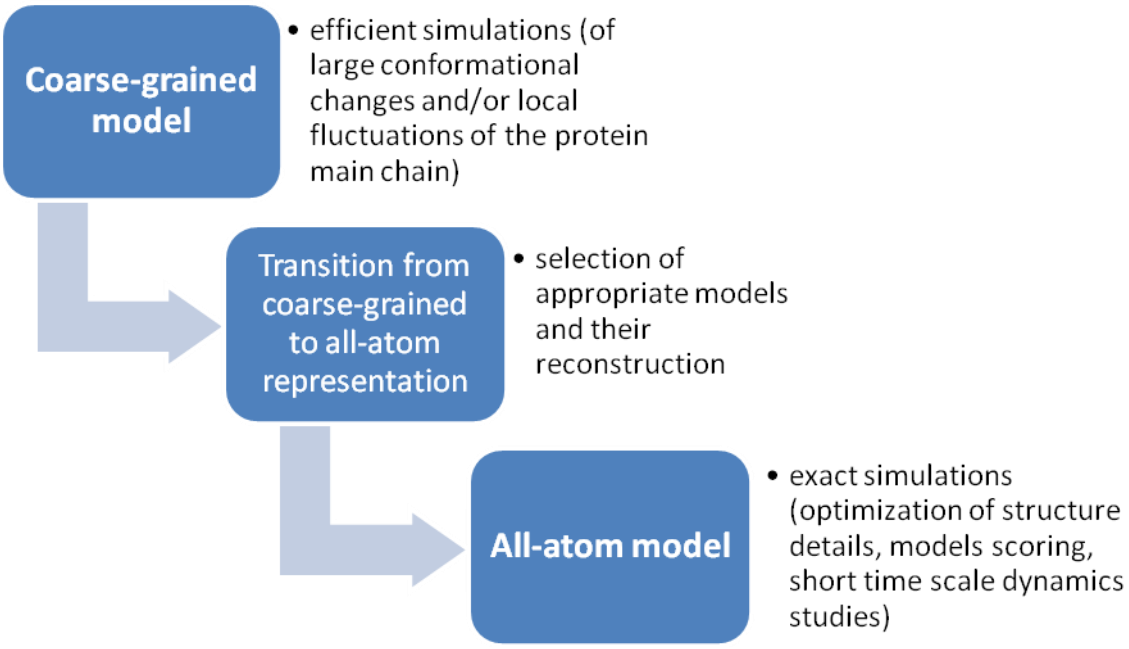

Figure 1. General pipeline of protein multiscale modeling used in the publications comprising the academic achievement.

Particular scientific goals and the presentation of the publications results

Particular scientific goals included:

I. Development of a method for modeling of structure fluctuations of globular proteins:
- (i) development of a coarse-grained method for prediction of protein structure fluctuations: comparison with results from all-atom molecular dynamics simulations **(publication H1)**;
- (ii) development of a multiscale method for prediction of protein structure fluctuations: server CABS-flex **(publication H2)**;
- (iii) comparison of CABS-flex predictions of protein flexibility with structure fluctuations observed in NMR ensembles **(publication H3).**

II. Development of methods for prediction of protein structure and dynamics:
- (i) development of a multiscale method for prediction of protein structure and dynamics: server CABS-fold **(publication H4)**;
- (ii) development of a multiscale method for modeling of extracellular loops in membrane G-coupled protein receptors **(publication H5)**.

III. Development of molecular modeling methods and their application in studies of the mechanisms of protein function:

(i) studies of the mechanism of chaperonin function **(publication H6)**;
(ii) studies of the folding mechanism of the B domain of protein A – transition state analysis **(publication H6 and H7)**;
(iii) studies of the mechanism of folding and binding of an intrinsically disordered protein **(publication H8)**.

IV. Development of protocols for multiscale protein modeling:
(i) development of a protocol for multiscale protein modeling with the use of fast optimization of all-atom models **(publication H7)**;
(ii) development of a protocol for multiscale protein modeling with the use of replica exchange molecular dynamics **(publication H9)**;
(iii) development of protocols for efficient simulations of long time protein dynamics **(publication H10)**.

V. Reviews of current literature, and own works, in the field of multiscale protein modeling:
(i) review article entitled „*Multiscale Approach to Protein Folding Dynamics"* **(publication H11)**;
(ii) review article entitled "*Coarse-Grained Modeling of Protein Dynamics"* **(publication H12)**.

In the last dozen years or so, all-atom molecular dynamics (MD) has become a popular tool used for simulation of protein structure fluctuations (near-native dynamics). Such simulations are most frequently performed by: (i) their combination with sparse experimental data about system dynamics, (ii) starting from a single, experimental native structure without experimental data guidance (Vendruscolo 2007; Fisette, Lague et al. 2012). In 2007, the common view that atomistic molecular dynamics captures the essential physics of protein near-native dynamics had been supported by the study of Orozco group (Rueda, Ferrer-Costa et al. 2007). Orozco showed that various MD force-fields provide a consistent picture of protein fluctuations in aqueous solution (10 nanosecond MD simulations, in explicit solvent, starting from native experimental structures). The simulations were performed for all protein metafolds using the four most popular force fields: OPLS, CHARMM, AMBER, and GROMOS, and required a massive supercomputer effort:

- computations took around 50 years of CPU time
- > 1.5 terabytes of trajectory data was obtained

In our study, we demonstrated that the aforementioned consensus view of protein dynamics from the Orozco work is fairly consistent with the dynamics of the coarse-grained protein model - the CABS model **(publication H1)**. Importantly, we roughly estimated CABS dynamics to be 6×10^3 cheaper in terms of computational cost than the all-atom MD. In our study, we performed optimization of CABS parameters (simulation time, temperature, various types of distance restraints) to obtain the best possible convergence with the MD data

from the Orozco study. The developed scheme of near-native simulations by the CABS model has been subsequently used as the fundamental element of an automated and multiscale method for fast simulation of protein structure fluctuations: the CABS-flex server **(publication H2)** (the server is available at http://biocomp.chem.uw.edu.pl/CABSflex). In the CABS-flex method, the following modeling stages were integrated: (i) simulations of near-native dynamics by the CABS model, (ii) CABS trajectory analysis and selection of representative models (using structural clustering), (iii) reconstruction of selected models to all-atom resolution and their further optimization – with the use of BBQ algorithm (Gront, Kmiecik et al. 2007) and, in the subsequent step, with the use of ModRefiner (Xu and Zhang 2011). The only data required as an input in the CABS-flex server is a protein structure file (given as a PDB code or uploaded by a user in PDB format). One of the CABS-flex server outputs is an ensemble of protein models (in all-atom PDB format) reflecting the flexibility of the input structure. The ensemble of predicted models can be used in structure-based studies of protein functions and interactions (e.g. in molecular docking or in analysis of the influence of near-native dynamics on the exposure of particular amino acid residues to the solvent[2]). Our subsequent studies focused on the comparison of protein fluctuations predicted by simulations (CABS-flex and all-atom MD) with protein structural variations within NMR ensembles **(publication H3)**. The comparison was done for a benchmark set of 140 proteins (determined by NMR and having MD simulation trajectories deposited in the MoDEL database (Meyer, D'Abramo et al. 2010)). The comparison has shown that the relative fluctuations of protein residues obtained from CABS-flex are well correlated with those derived from NMR ensembles. On average, this correlation is stronger than that between MD and NMR ensembles. Therefore, CABS-flex was shown to be an efficient alternative to conventional all-atom MD in predicting protein regions that undergo conformational changes as well as the extent of such changes **(publication H3)**.

As it was demonstrated in the studies of protein folding mechanisms (Kmiecik and Kolinski 2007; Kmiecik and Kolinski 2008), and in the protein structure prediction studies (Kolinski and Bujnicki 2005; Jamroz and Kolinski 2010), the CABS model makes possible the prediction of protein structure, or the structure of protein fragments of a relatively large length, based on sequence only, i.e. in a *de novo* fashion. *De novo* prediction is the most difficult modeling approach to protein structure prediction, possible only with the use of a small number of theoretical models. In order to provide an easy access to a unique de novo prediction CABS-based methodology, we developed an automatic and multiscale method for prediction of structure and dynamics of globular proteins: server CABS-fold **(publication H4)** (the server is available at http://biocomp.chem.uw.edu.pl/CABSfold).

Apart from de novo modeling procedures, the CABS-fold server has been equipped with procedures enabling modeling with the use of structural templates. In the CABS-fold server the following procedures have been integrated: (i) de novo modeling procedures, based on

[2] Such an analysis can be used in methods predicting interaction patches on protein surface (e.g. aggregation prone hydrophobic patches). Potential applications of the CABS-flex has been also summarized in the next part of this document.

sequence only, and modeling procedures that rely on multiple templates, with the use of the CABS model; (ii) trajectory analysis and selection of representative models with the use of structural clustering; (iii) reconstruction of selected models to all-atom resolution and their further optimization – with the use of BBQ algorithm (Gront, Kmiecik et al. 2007) and, in the subsequent step, with the use of ModRefiner (Xu and Zhang 2011). The output of the CABS-fold server includes a set of predicted models representing the most populated conformers in the CABS trajectory. Therefore, the predicted models can represent not only the near native conformations but also intermediate states on the path to the native conformation.[3]

As mentioned above, one of the unique features of the CABS model is the possibility of structure prediction of relatively large (in comparison to other algorithms) protein fragments based on sequence only. We've attempted to utilize this feature in the development of a method for structure prediction of the second extracellular loop in G-protein-coupled receptors (GPCRs) **(publication H5)**. GPCRs play key functions in living organisms. The second extracellular loop (ECL2) is a functionally important region of GPCRs which poses significant challenge for computational structure prediction methods. Therefore, the accurate prediction of ECL2 is critical for the construction of models applicable in drug design efforts (Peeters, van Westen et al. 2011; Wheatley, Wootten et al. 2012). We developed and tested several procedures of ECL2 structure modeling in thirteen GPCRs. The ECL2s (between 13 and 34 residue long) were predicted in an environment of other extracellular loops, which were fully flexible while the transmembrane domain was fixed in its X-ray conformation. Additionally, the modeling procedure utilized theoretical predictions of ECL2 secondary structure and experimental constraints on disulfide bridges. As the modeling result, we have obtained the following sets of models: (i) the lowest energy conformers (according to all-atom force-field) and (ii) the most populated conformers. The resulting sets of models contained models close to the available X-ray structures. The level of similarity between the predicted models and the X-ray structures was comparable to that of other state-of-the-art computational methods. Importantly, thanks to significant simplifications employed in our modeling scheme[4] the computational cost of the entire procedure was very low.

Considering the high flexibility of ECL2s (suggested but not precisely characterized by experiment) and its functional importance (Fraser 1989; Nanevicz, Wang et al. 1996; Zhao, Hwa et al. 1996; Ott, Troskie et al. 2002; Shi and Javitch 2002; Klco, Wiegand et al. 2005; Conner, Hawtin et al. 2007; Dror, Arlow et al. 2011; Wheatley, Wootten et al. 2012; Seibt, Schiedel et al. 2013) theoretical studies should aim at the characterization of an ensemble view of extracellular loops and its validation through experimental approaches. In our work **(publication H5)**, we presented an ensemble view of ECL2 structures (in sets of 10 or 100 cluster representatives or the lowest energy models). Analysis of the predicted ensembles suggests that, at least for some of the modeled receptors, ECL2s may be subjected to large

[3] Potential applications of the CABS-fold has been also summarized in the next part of this document.
[4] Coarse-grained CABS modeling that involves no membrane lipids, combined with a simple version of all-atom scoring and optimization.

molecular movements. However, the obtained sets of models were validated only by comparison with X-ray structures frozen in a single conformational option (unfortunately, there are no exact experimental data on ECL2 dynamics). In order to facilitate further experimental validation, and comparison with the results of other theoretical approaches, we made our models available at http://biocomp.chem.uw.edu.pl/GPCR-loop-modeling/.

Apart from simulation of protein folding, multiscale modeling procedures may also be used in simulations of even more complex processes, such as mechanisms of chaperone function **(publication H6)**. The so called chaperones are proteins responsible for assisting and enhancing the folding process of other proteins (Hartl and Hayer-Hartl 2009). The particular class of chaperones are chaperonins, a hollow cylindrical proteins that enable substrate proteins (SPs) to reach their native states by binding and encapsulating SPs inside the cylindrical structure. The exact mechanism of chaperonin assistance in protein folding remains unsolved. The observation of protein dynamics in the presence of or inside the chaperonin by means of experimental techniques is extremely difficult and lacks sufficient detail.

A number of models have been proposed for the chaperonin mode of action. These models are related to either of the postulated roles of the chaperonin cage: passive (aggregation prevention without altering the folding mechanism or kinetics) or active (folding promotion) (Hartl and Hayer-Hartl 2009; Jewett and Shea ; Lucent, England et al. 2009). Probably the earliest and most popular explanation of the chaperonin active role is the iterative annealing mechanism (IAM), which focuses on the effect of iterative interactions of SP with the hydrophobic chaperonin interior. In **publication H6**, we presented simulations of chaperonin-assisted protein folding with the use of the CABS model. The simulations were performed:

- for two small proteins, which are the paradigm systems of protein folding: B domain of protein A (BdpA) and B1 domain of protein G (GB1)
- in two versions: (1) chaperonin assisted simulations and (2) chaperonin-free simulations – in order to compare version (1) and (2)

So far, the simulation studies of the IAM suggested essentially the same mechanistic concept: binding to the hydrophobic walls of the open chaperonin molecule helps to pull out the kinetically trapped SPs from their misfolded conformations.(Lucent, England et al. 2009) We have demonstrated that periodical disrupting can also have an effect on a folding pathway modulation towards the more efficient exploration of the folding landscape and avoiding the kinetic traps. The iterative relaxation of the tertiary structure together with stabilization of the preformed secondary structure elements results in the folding mechanism shifting from nucleation-condensation towards the framework. Framework and nucleation-condensation mechanisms of protein folding are extreme manifestations of an underlying common mechanism.(Gianni, Guydosh et al. 2003; White, Gianni et al. 2005). The former is favored by a highly stable secondary structure, the latter is the concomitant secondary and tertiary

structure formation. A shift between the two extremes is not unusual upon condition changes or even due to a single mutation (Gianni, Guydosh et al. 2003; White, Gianni et al. 2005).

In the **H6 publication**, besides the chaperonin effect analysis, we presented also the folding mechanism studies of the BdpA protein. The prediction of the BdpA Transition State Ensemble (TSE - corresponding to the maximum of the free energy landscape) – in the form of contact maps and protein chain mobility analysis – is in perfect agreement with experimental findings (with phi value analysis). The characterization of structural features of the TSE had been further extended in **publication H7** by providing the structural analysis of the set of all-atom models.

Within our subsequent work, we have developed an efficient simulation method for folding and binding of an intrinsically disordered protein **(publication H8)**. The mechanism of these processes remains unknown. In the study, we used the complex of the phosphorylated kinase-inducible domain (pKID) with its interacting domain (KIX), which is a model system for studies of the mechanisms by which intrinsically unfolded proteins perform their functions. What is important, the simulations were performed without any prior knowledge about the experimental pKID structure in the bound form:

- simulations were started from the unbound, randomly positioned and disordered pKID structure
- during simulations the conformation of pKID was fully flexible (not restricted by any restraints) while the KIX dynamics was restricted to near-native fluctuations

Simulations of such large conformational changes, without any prior knowledge about the bound protein structure, remain inaccessible for classical simulation methods (and according to our knowledge haven not been performed before for the pKID/KIX system). In the simulation results, we obtained models of pKID in a near-native conformation and docked in the native binding site (see the movie showing an example trajectory: http://youtu.be/WhS8UFaEodk). The simulation analysis allowed us to obtain the characteristics of transient encounter complexes on the path to the native complex arrangement, which appeared to be in good agreement with experimental results (Sugase, Dyson et al. 2007). We have found that the key folding and binding step is linked to the formation of weak native interactions between a preformed native-like fragment of the pKID helix and the KIX surface. Once that nucleus forms, the pKID chain may condense from a largely disordered encounter ensemble to a natively bound and ordered conformation. The observed mechanism is reminiscent of a nucleation-condensation model, a common scenario for folding of globular proteins (Fersht 1995; Itzhaki, Otzen et al. 1995; Daggett and Fersht 2003). What's interesting, a similar concept of the binding and folding scenario has been recently proposed based on protein engineering and kinetic experiments (phi value analysis) of the ACTR/NCBD system (Dogan, Mu et al. 2013), or stopped-flow spectroscopy study of the reaction between disordered peptides and PDZ domains (Haq, Chi et al. 2012).

Apart from the development of multiscale methods suited to tackle the particular simulation problems (described above), we have also developed general-purpose procedures facilitating the merging of coarse-grained and all-atom modeling techniques. As a result of optimization and testing of various multiscale modeling procedures, we have developed:

- a protocol of multiscale modeling with the use of fast optimization of all-atom models **(publication H7)**
- a protocol of multiscale modeling with the use of replica exchange MD **(publication H9)**
- protocols for efficient simulations of long time protein dynamics using the coarse-grained CABS model – from the denatured to the native structure **(publication H10)**

The method described in **publication H7** allows for efficient and fast conversion of the coarse-grained trajectory (based on alpha-carbon trace only, e.g. from the CABS model) to the trajectory of all-atom models. The method presented in **publication H9** makes it possible to perform replica-exchange MD simulations in a more efficient way than all-atom simulations starting from the denatured chain conformation. This higher efficiency (acceleration of system convergence) was possible thanks to starting the simulations from the appropriate models taken from coarse-grained simulation results. The simulation methods described in **publication H10** were based on the strategies developed previously for simulations of the folding mechanisms: of barnase and chymotrypsin inhibitor (Kmiecik and Kolinski 2007) and B1 domain of protein G (Kmiecik and Kolinski 2008). The developed software can be easily used for simulation of new protein systems and analysis of the resulting data.

The results presented in the publications comprising the scientific achievement illustrate the unique capabilities of the developed methods in simulation of various macromolecular systems. The major advantage of the developed methods, in comparison to classical modeling techniques, is higher modeling efficiency which allowed for: (i) much faster performance in simulations of various tasks (which opens the possibility of carrying many simulation variants and for many systems); (ii) carrying simulations for processes, whose time scales are inaccessible to classical modeling methods.

Finally, the proposed molecular modeling methods offer the foundation to further development of the multiscale modeling approach, merging coarse-grained and all-atom modeling. Presently, it is expected that such direction of development will allow in the future for modeling of more complex macromolecular machines (e.g. large size protein complexes), which are in the center of interest of the biomedical research (see also review papers **H11**, **H12** and (Scheraga, Khalili et al. 2007; Russel, Lasker et al. 2009; Whitford, Noel et al. 2009)).

Discussion of the possible use of the obtained results

1. Utilization of the CABS-flex method

The CABS-flex method (the development and tests of which have been described in **publications H1, H2 and H3**) was made available as an easy to handle web server. During the last one and a half year (from May 2013 to September 2014), we noted 31.4 thousand visits on the server website and 5.9 thousand of unique server users. Till September 2014, the CABS-flex method has been used in various published studies concerning structure-function-dynamics relationship (of proteins that play important biological functions):

- performed by Nevo (Newo 2014) – the study was devoted to the analysis of structure dynamics of helicases (the analysis may find potential application in the design of anti-parasitic drugs targeting helicases)
- performed by Liu at al. (Liu, Werner et al. 2014) – the study was devoted to the analysis of structure dynamics of serpins (serine protease inhibitors) – CABS-flex was used as a supportive tool in the interpretation of experimental data
- performed by Fraga et al. (Fraga, Grana-Montes et al. 2014) – the study was devoted to the analysis of small protein domains rich in disulfide bonds – CABS-flex was used as a supportive tool in the interpretation of other theoretical predictions

The CABS-flex method was also tested as an element of the new strategy for prediction of aggregation hot-spots, developed in the Autonomic University of Barcelona, under the supervision of Professor Salvador Ventura (Pujagut 2013). The new strategy was named AGGRESCAN**3D**. The AGGRESCAN**3D** utilizes the following data:

- information about the protein structure (about the level of amino acid exposure to the solvent and about their mutual arrangement on the protein surface) and the information about the protein structure fluctuations (taken from the CABS-flex server),
- the propensity of a particular amino acid to aggregate – predicted by the AGGRESCAN method (this method is a popular and acknowledged tool for the prediction of aggregation "hot-spots" based on sequence only (Conchillo-Sole, de Groot et al. 2007; de Groot, Castillo et al. 2012)).

The tests of AGGRESCAN**3D** were performed in two versions: (1) with the use of an ensemble of models reflecting protein flexibility from the CABS-flex server, and (2) without using the information about protein flexibility. They have shown that utilization of flexibility predictions is necessary to obtain correct prediction results (Pujagut 2013). These studies are presently carried on within a collaboration between the Ventura group, myself and my coworkers. The cooperation goal is further validation and optimization of AGGRESCAN3D as well as making it available in the form of an easy to handle tool.

Apart from the mentioned above, the most promising applications of the CABS-flex method include its incorporation into various molecular modeling tasks:

- into methods predicting the tendency of proteins to aggregate – e.g., according to the scheme of the SAP (*spatial aggregation propensity*) technology developed by Chennamsetty et al. (Chennamsetty, Voynov et al. 2009). Chennamsetty et al., used all-atom MD to simulate the variability of amino acids exposure to the solvent (due to fluctuations of the structure in the solution). Replacing all-atom MD in the SAP scheme by the CABS-flex may potentially: (i) make the prediction much faster and more accurate; (ii) facilitate utilization of the SAP scheme for antibodies of large size (presently, it is not possible because of the computational cost of all-atom MD);
- in methods predicting interaction patches on the protein surface: (i) protein-protein interaction sites, (ii) protein – nucleic acid (DNA or RNA) binding sites, (iii) protein-ligand binding sites. Presently, most of the established methods performing task (i), (ii) or (iii), do not use the information about the variability of amino acid exposure to the solvent (which is involved in protein structure fluctuations). Thus, those methods can be easily extended by the utilization of the CABS-flex predictions, analogically to the above described AGGRESCAN**3D**;
- in molecular docking methods (according to the ensemble docking strategy (Korb, Olsson et al. 2012);
- in drug design or protein design pipelines e.g. in mutation or protein stability analyses applied to many protein variants (large screening analyses cannot be performed by classical all-atom MD because of too much computational cost);
- in theoretical methods facilitating interpretation of the experimental data about protein dynamics (e.g. in interpretation of NMR data, or in molecular replacement).

2. Utilization of the CABS-fold method

CABS-fold method **(publication H4)** was made available as an easy to handle web server. During the last one and a half years (from May 2013 to September 2014), we noted 15.0 thousand visits on the server website and 3.3 thousand of unique server users. Till September 2014, the CABS-fold method been used in various published studies concerning structure-function relationship (proteins that play important biological functions):

- performed by Cong et al. (Cong, Casiraghi et al. 2013) – the CABS-fold was applied in prediction of a prion protein
- performed by Mosalaganti (Mosalaganti 2014) – the CABS-fold was applied in prediction of a ROD protein fragment (ROD protein plays important role in the metaphase of the cell cycle)

In comparison to other structure prediction methods, the CABS model presents exceptional performance in prediction of protein structure from sequence only (so called *de novo*

modeling) for: (i) small proteins of the length up to 100 amino acids, or (ii) long protein fragments (Jamroz and Kolinski 2010). The unique capabilities of the CABS model have been also verified during 6th CASP (Critical Assessment of methods for Protein Structure Prediction) experiment (Kolinski and Bujnicki 2005). It is also worth to mention that CAS and TOUCHSTONE – methods based on CABS methodology by Zhang and coworkers (Zhang, Kolinski et al. 2003) – are currently one of the best methods for *de novo* structure prediction (Zhang, Kolinski et al. 2003; Roy, Kucukural et al. 2010; Xu, Zhang et al. 2011). De novo (and not only) modeling examples were also illustrated at CABS-fold web pages at http://biocomp.chem.uw.edu.pl/CABSfold/example_results.php.

Server CABS-fold is suited to perform the following modeling tasks:

- protein structure prediction based on sequence only (for each amino acid, a user can also specify the expected secondary structure type as well as distance restrains, which will be used during the simulation);
- protein structure prediction based on structural template(s) – in the case of many alternative templates a consensus solution is taken (consistent fragments of templates are considered as credible, while inconsistent ones are treated with a low confidence);
- structure prediction of protein loops or missing fragments – when a single structural template with a missing fragment (e.g. corresponding to loop) is provided as an input, the CABS-fold performs fragment modeling while the rest of the protein is nearly "frozen" in the template conformation;
- prediction of protein structure with the use of sparse experimental data. Such data can be introduced in the modeling process via the distance restraints between the chosen alpha-carbons;
- prediction of intermediate states during protein folding or prediction of transient structures for proteins which do not demonstrate a stable native structure (e.g. for disordered proteins) – in order to gain insight into the structure and dynamics of such proteins one should analyze all the predicted models (cluster representatives) and/or the coarse-grained simulation trajectory (available from the CABS-fold results).

3. Utilization of the other developed methods and the obtained results:
    - the developed method for prediction of the second extracellular loop in G-coupled protein receptors (**publication H5**) is an efficient alternative to other modeling approaches. Moreover, the results presented in work **H5** (and the predicted models that were made available within this work) constitute the benchmark for other theoretical studies (the modeling results were presented for 13 GPCRs and, for most of them, for the first time). The presented results suggest the possible improvements of the method;

- the results of studying the mechanism of chaperonin function (**publication H6**) gave rise to a new hypothesis, which can be tested by properly designed experimental studies. Moreover, in **publications H6** and **H7**, a detailed interpretation of the mechanism of BdpA folding had been presented, which can also be further tested using theory or experiment;
- a new simulation strategy of a coupled folding and binding mechanism of an intrinsically disordered protein, proposed in the **H8** work**,** qualitatively extended the possibilities of molecular docking and multiscale modeling of proteins. Namely, the performed simulations were unique in that they ensured full flexibility of the docked peptide, without–relying on any prior knowledge about the bound conformation of the docked peptide and the binding site. The obtained results facilitated interpretation of the available experimental data and allowed to formulate a general hypothesis concerning the mechanism of binding and folding of disordered proteins. Since there is a great need for effective simulation tools able to explore the functions of disordered proteins, we presently continue the development of the method presented in work **H8**. The next planned step is to optimize and accommodate the method in the form of an automatic web server
- the reconstruction procedure from coarse-grained to all-atom models and further optimization scheme in all-atom resolution, presented in work **H7**, has been already utilized in multiscale modeling studies (e.g. in publications **H5, H8** and (Kouza, Hu et al. 2013)), and is presently used as a basic framework for the development of new improved modeling methods carried out in the Laboratory of Theory of Biopolymers (Faculty of Chemistry, University of Warsaw). What is important, the procedure is based on the publicly available software, therefore it can be freely used and modified according to other users' needs (also in combination with coarse-grained models different than the CABS model).
- the multiscale modeling procedure presented in work **H9** is presently used in the studies carried out in the Laboratory of Theory of Biopolymers (Faculty of Chemistry, University of Warsaw) – in application to simulations of systems larger than that studied in the **H9** work and in the development of new, improved modeling strategies
- protocols for efficient simulations of long time protein dynamics using the coarse-grained CABS model, presented in publication **H10**, allows for automated running of a broad set of coarse-grained simulations, and their further analysis according to schemes introduced and tested in the following works: (Kmiecik and Kolinski 2007; Kmiecik and Kolinski 2008).

In the publications comprising the scientific achievement, we proposed publicly available and easy-to-handle tools for molecular modeling: server CABS-flex (**publications H1**, **H2**, **H3**), and server CABS-fold (**publication H4**). Such servers are very useful in popularization of new computational methods and, what seems to be extremely important, have made these advanced methods easily available for experimental biologists, chemists and

other experts in life science. Moreover, the presented publications describe innovative tools and strategies of multiscale modeling (**H7**, **H8**, **H9**, **H10**), the elements of which are available for download (from the laboratory website http://biocomp.chem.uw.edu.pl/tools or from other sources – if the tools developed by other research groups were used). Apart from molecular modeling tools, we provided also a large collection of various types of data, including 3D models, which can be utilized or verified in theoretical or experimental studies of the studied molecular mechanisms (**publications H5**, **H6**, **H7**, **H8**).

References

Chennamsetty, N., V. Voynov, et al. (2009). "Design of therapeutic proteins with enhanced stability." Proc Natl Acad Sci U S A **106**(29): 11937-11942.

Chouard, T. (2011). "Structural biology: Breaking the protein rules." Nature **471**(7337): 151-153.

Conchillo-Sole, O., N. S. de Groot, et al. (2007). "AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides." Bmc Bioinformatics **8**: 65.

Cong, X., N. Casiraghi, et al. (2013). "Role of Prion Disease-Linked Mutations in the Intrinsically Disordered N-Terminal Domain of the Prion Protein." Journal of Chemical Theory and Computation **9**(11): 5158-5167.

Conner, M., S. R. Hawtin, et al. (2007). "Systematic analysis of the entire second extracellular loop of the V(1a) vasopressin receptor: key residues, conserved throughout a G-protein-coupled receptor family, identified." J Biol Chem **282**(24): 17405-17412.

Daggett, V. and A. R. Fersht (2003). "Is there a unifying mechanism for protein folding?" Trends Biochem Sci **28**(1): 18-25.

de Groot, N. S., V. Castillo, et al. (2012). "AGGRESCAN: method, application, and perspectives for drug design." Methods Mol Biol **819**: 199-220.

Dogan, J., X. Mu, et al. (2013). "The transition state structure for coupled binding and folding of disordered protein domains." Sci Rep **3**: 2076.

Dror, R. O., D. H. Arlow, et al. (2011). "Activation mechanism of the beta2-adrenergic receptor." Proc Natl Acad Sci U S A **108**(46): 18684-18689.

Fersht, A. R. (1995). "Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications." Proc Natl Acad Sci U S A **92**(24): 10869-10873.

Fisette, O., P. Lague, et al. (2012). "Synergistic applications of MD and NMR for the study of biological systems." J Biomed Biotechnol **2012**: 254208.

Fraga, H., R. Grana-Montes, et al. (2014). "Association Between Foldability and Aggregation Propensity in Small Disulfide-Rich Proteins." Antioxidants & Redox Signaling **21**(3): 368-383.

Fraser, C. M. (1989). "Site-directed mutagenesis of beta-adrenergic receptors. Identification of conserved cysteine residues that independently affect ligand binding and receptor activation." J Biol Chem **264**(16): 9266-9270.

Gianni, S., N. R. Guydosh, et al. (2003). "Unifying features in protein-folding mechanisms." Proc Natl Acad Sci U S A **100**(23): 13286-13291.

Gront, D., S. Kmiecik, et al. (2007). "Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates." J Comput Chem **28**(9): 1593-1597.

Haq, S. R., C. N. Chi, et al. (2012). "Side-chain interactions form late and cooperatively in the binding reaction between disordered peptides and PDZ domains." Journal of the American Chemical Society **134**(1): 599-605.

Hartl, F. U. and M. Hayer-Hartl (2009). "Converging concepts of protein folding in vitro and in vivo." Nat Struct Mol Biol **16**(6): 574-581.

Henzler-Wildman, K. A., M. Lei, et al. (2007). "A hierarchy of timescales in protein dynamics is linked to enzyme catalysis." Nature **450**(7171): 913-916.

Itzhaki, L. S., D. E. Otzen, et al. (1995). "The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding." J Mol Biol **254**(2): 260-288.

Jamroz, M. and A. Kolinski (2010). "Modeling of loops in proteins: a multi-method approach." BMC Struct Biol **10**: 5.
Jewett, A. I. and J. E. Shea (2009). "Reconciling theories of chaperonin accelerated folding with experimental evidence." Cell Mol Life Sci **67**(2): 255-276.
Karplus, M. and J. Kuriyan (2005). "Molecular dynamics and protein function." Proceedings of the National Academy of Sciences of the United States of America **102**(19): 6679-6685.
Klco, J. M., C. B. Wiegand, et al. (2005). "Essential role for the second extracellular loop in C5a receptor activation." Nat Struct Mol Biol **12**(4): 320-326.
Kmiecik, S. and A. Kolinski (2007). "Characterization of protein-folding pathways by reduced-space modeling." Proc Natl Acad Sci U S A **104**(30): 12330-12335.
Kmiecik, S. and A. Kolinski (2008). "Folding pathway of the b1 domain of protein G explored by multiscale modeling." Biophysical Journal **94**(3): 726-736.
Kolinski, A. and J. M. Bujnicki (2005). "Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models." Proteins-Structure Function and Bioinformatics **61 Suppl 7**: 84-90.
Korb, O., T. S. G. Olsson, et al. (2012). "Potential and Limitations of Ensemble Docking." J Chem Inf Model **52**(5): 1262-1274.
Kouza, M., C.-K. Hu, et al. (2013). "A structure-based model fails to probe the mechanical unfolding pathways of the titin I27 domain." J Chem Phys **139**(6): -.
Levitt, M. and A. Warshel (1975). "Computer-Simulation of Protein Folding." Nature **253**(5494): 694-698.
Liu, L., M. Werner, et al. (2014). "Collapse of a Long Axis: Single-Molecule Förster Resonance Energy Transfer and Serpin Equilibrium Unfolding." Biochemistry **53**(18): 2903-2914.
Liwo, A., Y. He, et al. (2011). "Coarse-grained force field: general folding theory." Phys Chem Chem Phys **13**(38): 16890-16901.
Lucent, D., J. England, et al. (2009). "Inside the chaperonin toolbox: theoretical and computational models for chaperonin mechanism." Phys Biol **6**(1): 015003.
Meyer, T., M. D'Abramo, et al. (2010). "MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories." Structure **18**(11): 1399-1409.
Mosalaganti, S. (2014). Structural studies on human RZZ and Cln3p. PhD, Max-Planck-Institut für Molekulare Physiologie.
Nanevicz, T., L. Wang, et al. (1996). "Thrombin receptor activating mutations. Alteration of an extracellular agonist recognition domain causes constitutive signaling." J Biol Chem **271**(2): 702-706.
Newo, A. S. (2014). "Molecular Modeling of the Plasmodium falciparum Pre-mRNA Splicing and Nuclear Export Factor PfU52." The Protein Journal **33**(4): 354-368.
Ott, T. R., B. E. Troskie, et al. (2002). "Two mutations in extracellular loop 2 of the human GnRH receptor convert an antagonist to an agonist." Mol Endocrinol **16**(5): 1079-1088.
Peeters, M. C., G. J. van Westen, et al. (2011). "Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation." Trends Pharmacol Sci **32**(1): 35-42.
Pujagut, R. (2013). AGGRESCAN3D: A tool to predict aggregation propensity in protein surfaces MSc, Univestitat Autònoma de Barcelona.
Roy, A., A. Kucukural, et al. (2010). "I-TASSER: a unified platform for automated protein structure and function prediction." Nat Protoc **5**(4): 725-738.
Rueda, M., C. Ferrer-Costa, et al. (2007). "A consensus view of protein dynamics." Proc Natl Acad Sci U S A **104**(3): 796-801.
Russel, D., K. Lasker, et al. (2009). "The structural dynamics of macromolecular processes." Curr Opin Cell Biol **21**(1): 97-108.
Scheraga, H. A., M. Khalili, et al. (2007). "Protein-folding dynamics: overview of molecular simulation techniques." Annu Rev Phys Chem **58**: 57-83.

Seibt, B. F., A. C. Schiedel, et al. (2013). "The second extracellular loop of GPCRs determines subtype-selectivity and controls efficacy as evidenced by loop exchange study at A2 adenosine receptors." Biochem Pharmacol **85**(9): 1317-1329.
Shaw, D. E. (2013). "Millisecond-Long Molecular Dynamics Simulations of Proteins on a Special-Purpose Machine." Biophysical Journal **104**(2): 45a-45a.
Shaw, D. E., P. Maragakis, et al. (2010). "Atomic-Level Characterization of the Structural Dynamics of Proteins." Science **330**(6002): 341-346.
Shi, L. and J. A. Javitch (2002). "The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop." Annu Rev Pharmacol Toxicol **42**: 437-467.
Sugase, K., H. J. Dyson, et al. (2007). "Mechanism of coupled folding and binding of an intrinsically disordered protein." Nature **447**(7147): 1021-1025.
Takada, S. (2012). "Coarse-grained molecular simulations of large biomolecules." Curr Opin Struct Biol **22**(2): 130-137.
Vendruscolo, M. (2007). "Determination of conformationally heterogeneous states of proteins." Curr Opin Struct Biol **17**(1): 15-20.
Vendruscolo, M. and C. M. Dobson (2011). "Protein Dynamics: Moore's Law in Molecular Biology." Current Biology **21**(2): R68-R70.
Warshel, A. and M. Levitt (1976). "Theoretical Studies of Enzymic Reactions - Dielectric, Electrostatic and Steric Stabilization of Carbonium-Ion in Reaction of Lysozyme." J Mol Biol **103**(2): 227-249.
Wheatley, M., D. Wootten, et al. (2012). "Lifting the lid on GPCRs: the role of extracellular loops." Br J Pharmacol **165**(6): 1688-1703.
White, G. W., S. Gianni, et al. (2005). "Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding." J Mol Biol **350**(4): 757-775.
Whitford, P. C., J. K. Noel, et al. (2009). "An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields." Proteins **75**(2): 430-441.
Xu, D., J. Zhang, et al. (2011). "Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement." Proteins-Structure Function and Bioinformatics **79 Suppl 10**: 147-160.
Xu, D. and Y. Zhang (2011). "Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization." Biophysical Journal **101**(10): 2525-2534.
Zhang, Y., A. Kolinski, et al. (2003). "TOUCHSTONE II: a new approach to ab initio protein structure prediction." Biophysical Journal **85**(2): 1145-1164.
Zhao, M. M., J. Hwa, et al. (1996). "Identification of critical extracellular loop residues involved in alpha 1-adrenergic receptor subtype-selective antagonist binding." Mol Pharmacol **50**(5): 1118-1126.

Sebastian Kmiecik