Osiągnięcia Nauki i Techniki
Kierunki Rozwoju i Metody
KONWERSATORIUM POLITECHNIKI WARSZAWSKIEJ
Wkładka nr 10 do Miesięcznika Politechniki Warszawskiej nr 2/2007
Redaktor merytoryczny — Stanisław Janeczko

# Wieloskalowe modelowanie molekularne białek

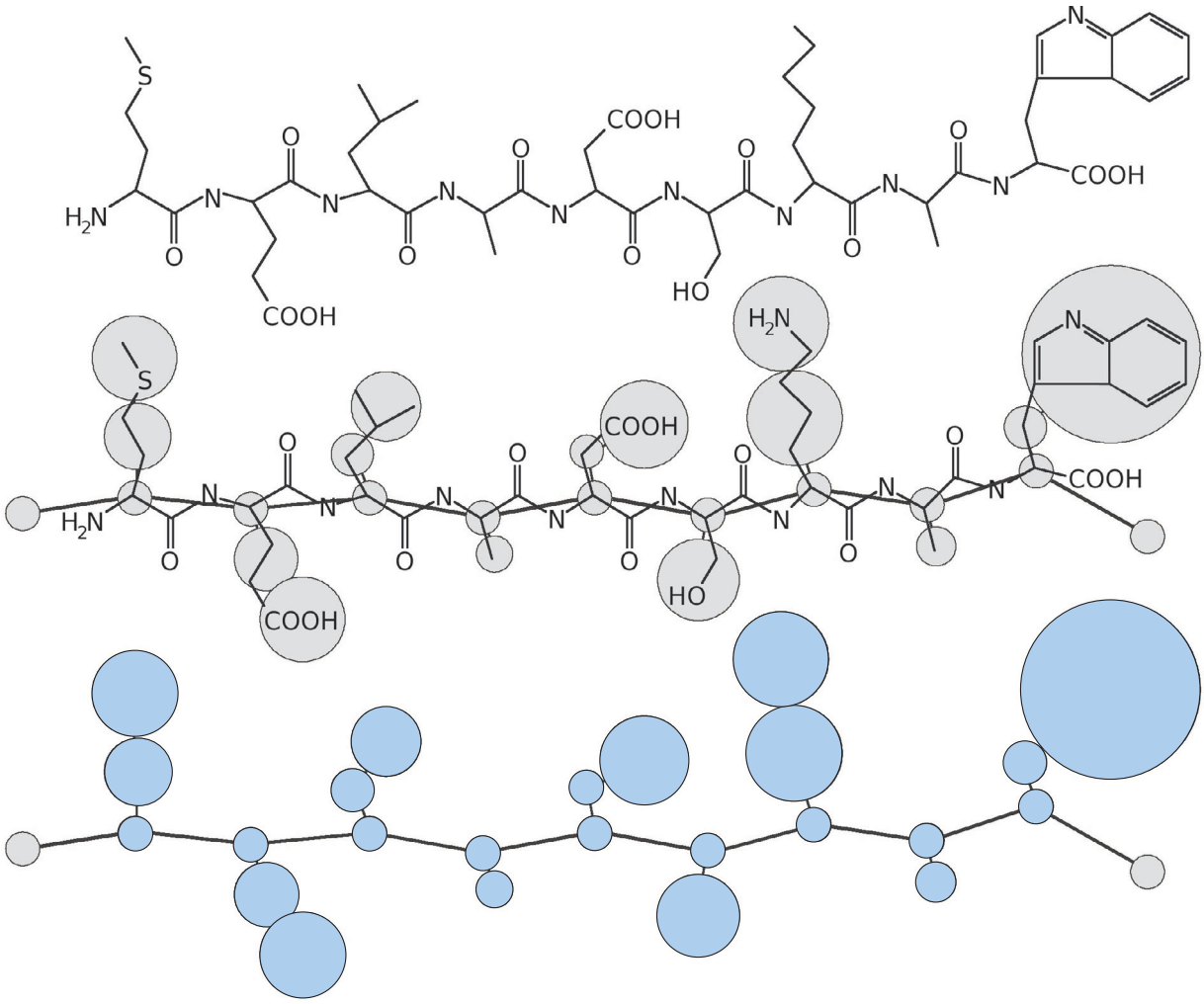
Na podstawie odczytu wygłoszonego w dniu 16 listopada 2006 roku

**Andrzej Koliński**

Pracownia Teorii Biopolimerów
Wydział Chemii, Uniwersytet Warszawski
e-mail: kolinski@chem.uw.edu.pl

Znamy obecnie około 30 milionów sekwencji aminokwasowych białek. Struktury przestrzenne udało się wyznaczyć doświadczalnie jedynie dla około 30 tysięcy z nich. Ta dysproporcja stale się powiększa. Sekwencjonowanie materiału genetycznego i tłumaczenie wyników na sekwencje aminokwasów jest tanie i wykonywane na ogół przez automaty, natomiast wyznaczenie ich struktury — czasochłonne, drogie i wymaga udziału wysokiej klasy specjalistów. Znajomość struktur jest niezbędna dla wielu celów [1], np. projektowania nowych leków i biotechnologii, zrozumienia molekularnych mechanizmów procesów chorobotwórczych itp., stąd olbrzymie znaczenie teoretycznych metod przewidywania struktury białek.

Procesy molekularne w organizmach żywych są zwykle związane z bardzo skomplikowanym przegrupowaniem olbrzymiej liczby atomów i molekuł, a charakterystyczne czasy tych przemian (czy reakcji) mogą różnić się między sobą o rzędy wielkości. Dla przykładu, proces spontanicznego zwijania się białek globularnych od losowej struktury zdenaturowanej do mniej lub bardziej jednoznacznej globularnej struktury natywnej trwa od milisekund do minut, zależnie od wielkości molekuły, typu struktury, warunków zewnętrznych itd. Znane są jednak dość liczne wyjątki dużo szybszych i dużo wolniejszych procesów denaturacji-reanaturacji. Tylko bardzo szybkie (od pikosekund do mikrosekund), a co z tym na ogół się wiąże bardzo lokalne procesy, można dziś modelować metodami klasycznej mechaniki molekularnej. W ciągu ostatnich 10–15 lat pokazano, że dobrze zaprojektowane zredukowane modele makromolekuł mogą być bardzo przydatnymi narzędziami modelowania molekularnego dużej (w odniesieniu do liczby atomów i czasu trwania procesu) skali [2–3]. Jeden ze starszych algorytmów autora, służących do mezoskopowego modelowania białek metodą Monte Carlo, został zintegrowany z pakietem dynamiki molekularnej CHARMM i jest publicznie dostępny na stronach internetowych The Scripps Research Institute (MMTSB — *Multiscale Modeling Tools for Structural Biology*. http://mmtsb.scripps.edu/software/mmtsbToolSet.html). Jest to pierwszy przykład ogólnie dostępnej i w pełni zautomatyzowanej metody do modelowania dynamiki białek na różnych poziomach rozdzielczości. W najbar-



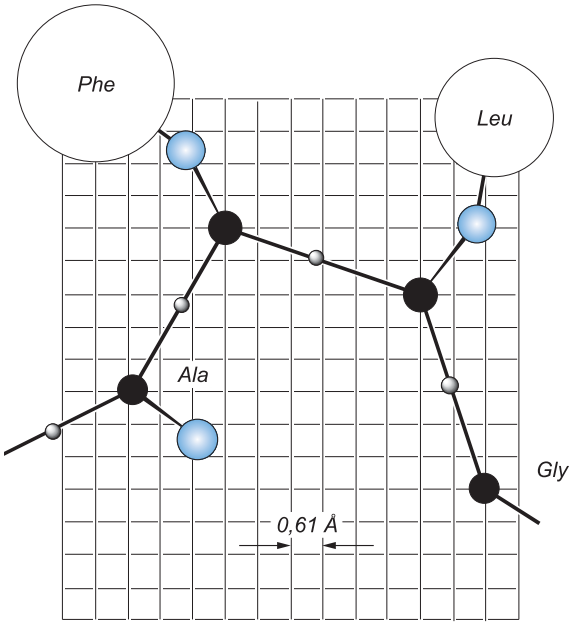

Rysunek 1. **Schemat redukcji reprezentacji geometrycznej krótkiego fragmentu łańcucha polipeptydowego. Górny panel pokazuje reprezentację pełnoatomową, w której w celu zwiększenia czytelności pominięto atomy wodoru i zaznaczono odpowiednimi literami nazwy niektórych atomów. Środkowy panel pokazuje sposób łączenia grup atomów w formie „zjednoczonych atomów", dolny panel — otrzymaną pseudostrukturę — zredukowany model fragmentu polipeptydowego**

dziej ogólnym sformułowaniu zasada działania MMTSB w zastosowaniu do dynamiki białek polega na wykonaniu symulacji za pomocą przybliżonego (ale za to bardzo szybkiego) modelu niskiej rozdzielczości (na poziomie zjednoczonych reszt), odbudowaniu detali atomowych dla wybranych klatek trajektorii niskiej rozdzielczości i wykonaniu szczegółowej analizy dynamiki pełnoatomowej CHARMM dla krótkich przedziałów czasu.

Zredukowane modele molekularne białek projektowane są w różnorodny sposób [2], ale te najbardziej sprawdzone w praktyce zakładają zwykle uproszczoną reprezentację łańcucha głównego w postaci łańcucha pseudowiązań pomiędzy atomami węgli alfa. Łańcuchy boczne aminokwasów zastępowane są zazwyczaj jednym lub dwoma zjednoczonymi atomami, reprezentującymi fragmenty grup bocznych. Przykładowy sposób redukcji geometrycznej reprezentacji łańcucha polipeptydowego pokazano na rysunku 1.

Model zredukowany CABS rozwijany przez autora [3] zakłada podobny do naszkicowanego powyżej sposób reprezentacji polipeptydów. Łańcuch główny ograniczony jest do szkieletu węgli alfa. Dodatkowo położenia węgli alfa ograniczone są do węzłów prostej sieci kubicznej o skoku 0,61 Å, co odpowiada około 1/2 średniej długości wiązań atomowych. Wykorzystanie reprezentacji siatkowej ma istotne znaczenie praktyczne, pozwalając wielokrotnie przyspieszyć procesy obliczeniowe w porównaniu z równoważnymi modelami w ciągłej przestrzeni stanów. Grupy boczne reprezentowane są za pomocą dwóch zjednoczonych atomów — jednego centrowanego na węglu beta, a drugiego w środku masy pozostałej części grupy bocznej, o ile aminokwas takie posiada. Model zawiera dodatkowy pseudoatom umieszczony na środku wirtualnego wiązania między węglami alfa. Ten pseudoatom wykorzystywany jest do definicji kierunkowych oddziaływań naśladujących wiązania wodorowe między grupami peptydowymi łańcucha głównego. Na rysunku 2 przedstawiono w sposób schematyczny geometrię modelu CABS.

Oddziaływania molekularne w modelu CABS reprezentowane są przez szereg potencjałów średniej siły wyprowadzonych na podstawie statystycznej analizy




Rysunek 2. **Geometria modelu CABS (skrót od C-Alpha, Beta, Side-group). Czarne kulki symbolizują położenie węgla alfa, ograniczone do siatki. Liczba możliwych pseudowiązań (wektorów siatkowych) Ca-Ca wynosi 800, eliminując w ten sposób ewentualne artefakty sieci. Na środkach wiązań Ca-Ca zaznaczono pseudoatomy używane do obliczania oddziaływań naśladujących wiązania wodorowe. Szkielet Ca definiuje także dogodny układ współrzędnych lokalnych do obliczania położeń węgli beta (jaśniejsze kulki) zjednoczonych atomów zastępujących pozostałe części grup bocznych (duże białe kule). Rozmiary kulek na rysunku nie odpowiadają rzeczywistym rozmiarom odpowiednich grup atomów**

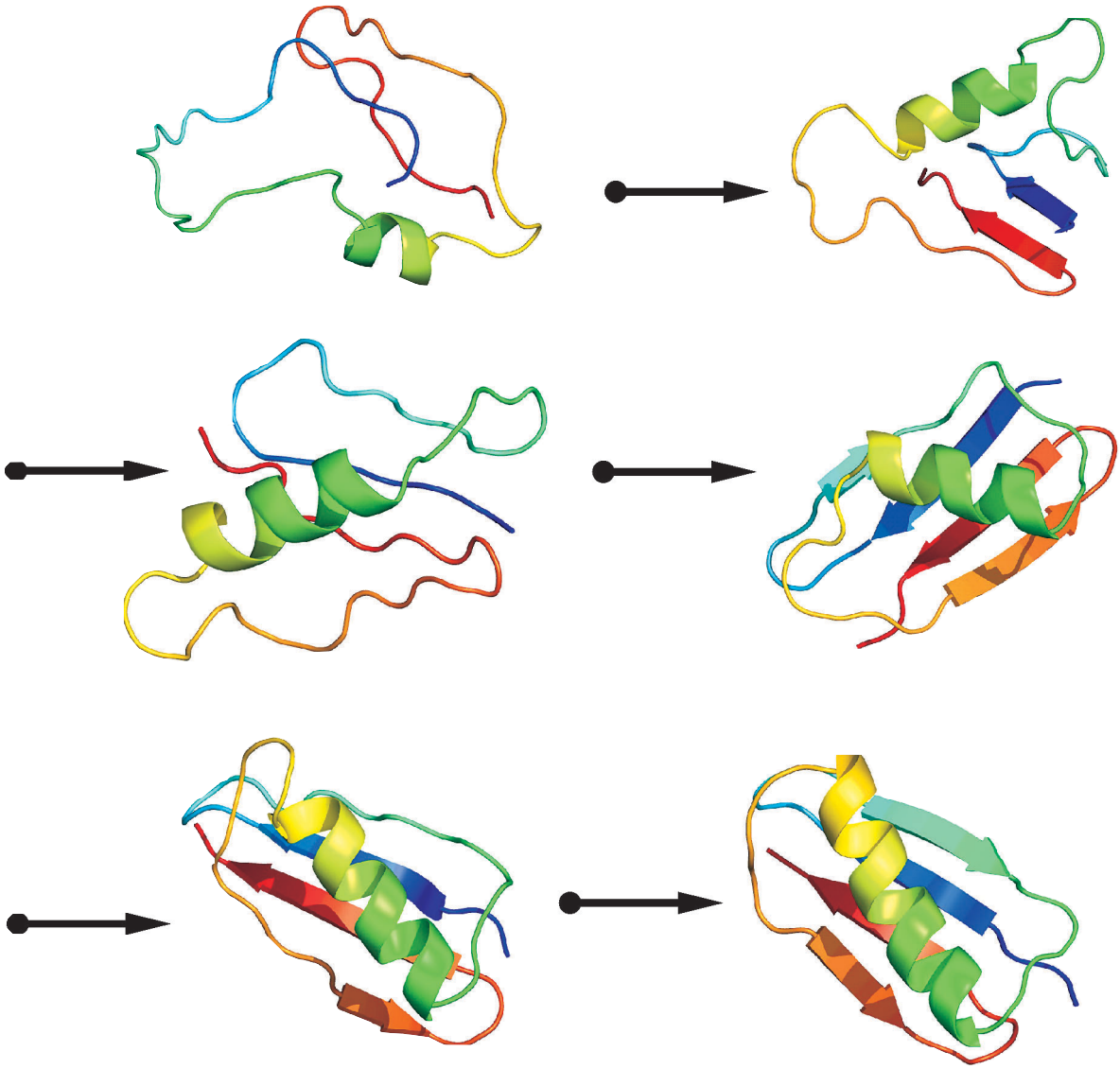
regularności strukturalnych obserwowanych w już znanych strukturach białek. Potencjały te odzwierciedlają tendencje odpowiednich sekwencji aminokwasów do przyjmowania określonej lokalnej geometrii łańcucha, którą z kolei określają odpowiednie preferencje co do struktury drugorzędowej (helis, beta-kartek itd.). Podobnie potencjały kontaktowe dla grup bocznych odzwierciedlają preferencje do charakterystycznego dla białek upakowania tych grup w stanie natywnym. Model oddziaływań CABS różni się w sposób jakościowy od innych mezoskopowych modeli białek. Między innymi wprowadzono tu jawną zależność różnych oddziaływań od lokalnej geometrii łańcucha, uwzględniając w ten sposób skomplikowane efekty wielociałowe, szczególnie ważne dla modeli zredukowanych. Szczegółowy opis pola sił modelu CABS można znaleźć w niedawno opublikowanym artykule [3] oraz na stronach internetowych autora (http://www.biocomp.chem.uw.edu.pl).

Dyskretna przestrzeń konformacyjna łańcuchów polipeptydowych w reprezentacji CABS próbkowana jest za pomocą metody Monte Carlo. Proces obliczeniowy polega na wykonaniu olbrzymiej liczby małych losowych zmian konformacji łańcucha w jego losowo wybranych punktach [1–2]. Zmiany takie akceptowane są zgodnie ze znanym kryterium Metropolisa. Lokalne mikromodyfikacje zaprojektowane są w sposób pozwalający interpretować ich długą sekwencję jako model dynamiki stochastycznej łańcucha w roztworze. Wykorzystywane są różne schematy symulacji: izotermiczne, tzw. symulowane schładzanie i metoda wymiany replik Monte Carlo [2].

Zredukowana reprezentacja przestrzeni konformacyjnej CABS i szybkie metody próbkowania pozwalają symulować *de novo* (tzn. wychodząc z sekwencji aminokwasów jako jedynej informacji specyficznej dla rozważanego białka) cały proces zwijania się małych białek globularnych. Na rysunku 3 pokazano 6 fotografii wybranych z trajektorii symulowanego schładzania małego białka globularnego — domeny B1 białka G (symbol 2GB1 w Protein Data Bank, PDB). Końcowa klatka przedstawia strukturę bardzo podobną do struktury natywnej — 1,9 Å RMSD (*Root-Mean-Square Deviation* — średnia kwadratowa odległość odpowiadających sobie atomów po najlepszym wzajemnym nałożeniu obu struktur, modelowej i doświadczalnej, z PDB). Wykonanie takiej symulacji zajmuje kilkanaście minut do godziny czasu pracy pojedynczego procesora PC.

CABS jest elementem centralnym szeregu wieloskalowych procedur do przewidywania struktury białek i asocjatów białkowych, modelowania molekularnego białek na podstawie fragmentarycznych danych doświadczalnych, czy też obliczeniowych badań termodynamiki i dynamiki białek. Najbardziej ogólnie, wieloskalowość modelowania polega na wykonaniu odpowiednich symulacji za pomocą algorytmu CABS i wybraniu odpowiednich struktur (czy sekwencji struktur) w reprezentacji zredukowanej, które następnie stanowią punkt wyjścia dla bardziej szczegółowego modelowania na poziomie atomowym.

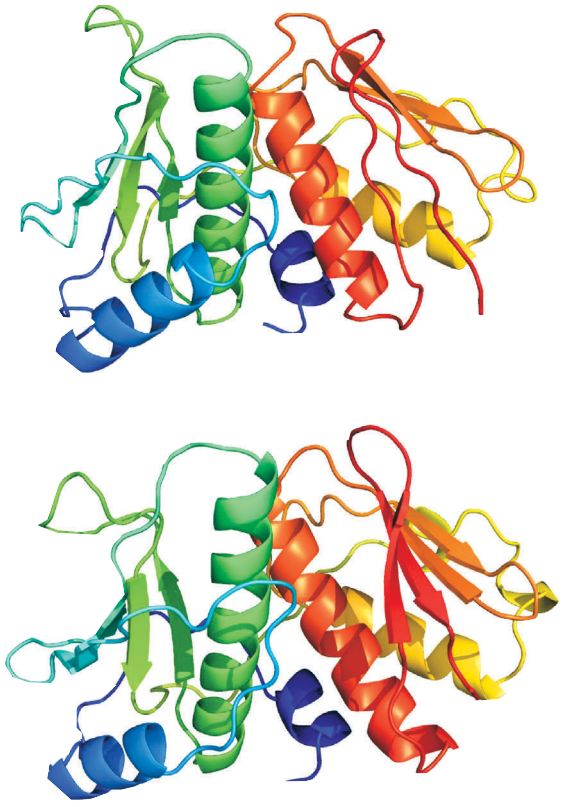
Nieco inne podejście wieloskalowe oparte na modelu CABS zostało zastosowane podczas eksperymentu CASP6 (*6$^{th}$ Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction* — CASP6), w którym wzięło udział ponad 200 grup z całego świata. Eksperyment CASP polega na komputerowym modelowaniu struktur białek, które właśnie mają być wyznaczone doświadczalnie w bliskiej przyszłości. Po opublikowaniu struktur doświadczalnych ocenia się zgodność wcześniej zdeponowanych w bezpiecznych komputerach organizatorów CASP modeli obliczeniowych (http://predictioncenter.org/casp6). Procedura modelowania zastosowana podczas CASP6 przez grupę Koliński-Bujnicki [4] polegała na zbudowaniu szeregu, często bardzo przybliżonych i niepełnych, modeli na podstawie ogólnodostępnych serwerów bioinformatycznych. Modele te służyły do zgromadzenia dużej liczby więzów odległości pomiędzy atomami modelowanej struktury. Więzy użyto w algorytmie CABS do ograniczenia, w przybliżony sposób, przestrzeni konformacyjnej modelu. Dużą liczbę otrzymanych modeli w zredukowanej przestrzeni konformacyjnej CABS poddano następnie analizie skupień [5-6], odbudowie de-



Rysunek 3. **Wybrane klatki trajektorii z typowej symulacji procesu zwijania struktury małego białka globularnego. Pierwsza klatka odpowiada startowej strukturze kłębka losowego, a ostatnia przedstawia strukturę końcową — bardzo podobną do struktury natywnej. Dla klarowności rysunku wybrano reprezentację wstęgową łańcucha polipeptydowego, pomijając szczegóły atomowe**

tali atomowych [7] dla reprezentatywnych struktur, a wreszcie ich ocenie i porządkowaniu na poziomie reprezentacji pełnoatomowej. Taka hierarchiczna procedura okazała się bardzo wydajna — grupa Koliński-Bujnicki została sklasyfikowana jako druga pod względem poprawności przewidzianych modeli molekularnych. Przykład struktury teoretycznie przewidzianej podczas CASP6 [4] pokazano na rysunku 4.

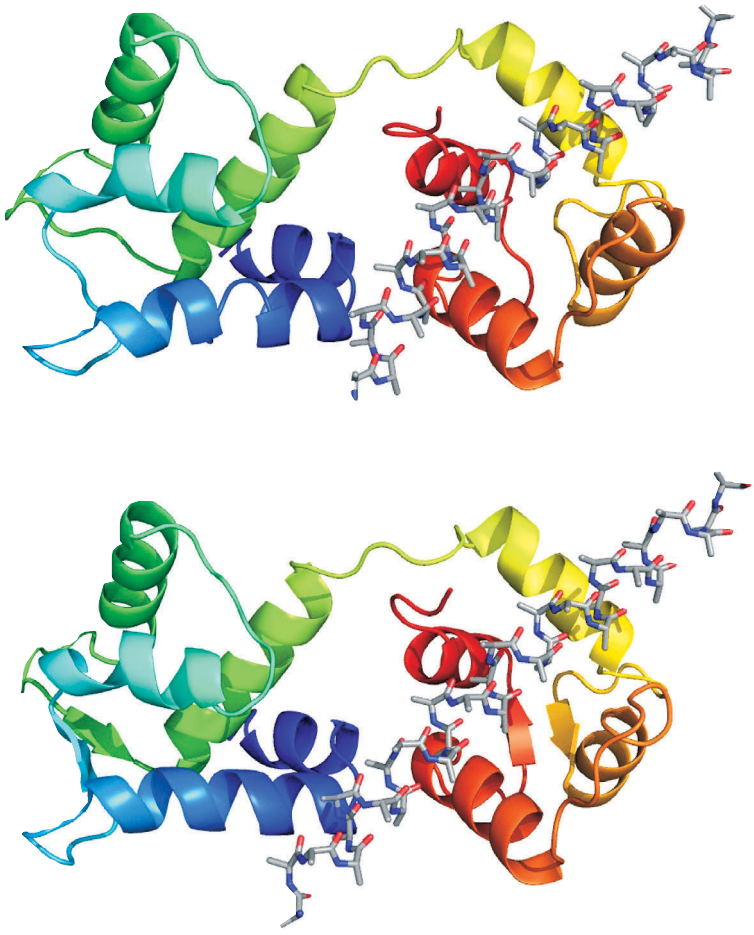
Podejście wieloskalowe pozwala również modelować układy składające się z większej liczby biomakromolekuł. Ma to znaczenie dla wspomaganego komputerowo racjonalnego projektowania nowych leków [8], zrozumienia procesów regulacyjnych w żywych komórkach, czy wyjaśnienia molekularnych podstaw niektórych procesów chorobowych [9]. Bardzo ważnym zagadnieniem molekularnej biologii obliczeniowej jest tzw. dokowanie. Dokowanie polega na znalezieniu sposobu łączenia się dwóch lub więcej molekuł i określeniu struktury powstałego kompleksu. Typowe zadanie obliczeniowe często wykonywane podczas projektowania leków [8] polega na zadokowaniu związku małocząsteczkowego (liganda) do znanej struktury białka (receptora). O ile znana jest struktura receptora, tradycyjne metody mechaniki molekularnej pozwalają czasami przewidzieć konformację liganda i miejsce jego dokowania. Słaba strona tego podejścia wynika z faktu, że struktura receptora w kompleksie może jakościowo różnić od jego struktury w izolacji lub w innym kompleksie. Tzw. „giętkie dokowanie" w tradycyjnej (pełnoatomowej) mechanice molekularnej ogranicza się do optymalizacji jedynie małych fragmentów receptora. Zastosowanie zredukowanego modelu CABS umożliwia traktowanie w sposób giętki całego kompleksu [10]. Obecnie pole siłowe CABS ograniczone jest do białek i peptydów. Pla-



Rysunek 4. **Przewidziana teoretycznie (góra rysunku) i doświadczalna (dół rysunku) struktura celu T0223. W celu schematycznego przedstawienia struktur użyto modelu wstęgowego opartego na węglach alfa. Kolor niebieski oznacza N-końce, a czerwony C-końce łańcuchów polipeptydowych. Model różni się od struktury krystalograficznej o 3 Å RMSD. Głównym źródłem błędu są dwa krótkie kawałki pętli na powierzchni N-końcowej domeny i między domenami (obie w odcieniach zieleni). Poprawnie przewidziano również wzajemną orientację domen**

nowane jest rozszerzenie umożliwiające modelowanie również ligandów niepeptydowych. W pokazanych tu przykładach dokowania wieloskalowego założono ograniczoną swobodę konformacyjną całego receptora i nieograniczoną swobodę liganda (peptydu lub drugiej molekuły białka). Startowa struktura receptora może pochodzić z badań doświadczalnych lub być wynikiem modelowania porównawczego (homologicznego). Symulacje CABS prowadzono metodą wymiany replik Monte Carlo, a końcowe struktury optymalizowano za pomocą pełnoatomowej mechaniki molekularnej [10], po uprzednim odbudowaniu reprezentacji atomowej. Struktury modelowych kompleksów porównano z ich strukturami wyznaczonymi doświadczalnie. Modelowano bardzo różne struktury, gdzie większa molekuła składała się z 31–179 aminokwasów, a mniejsza z 5–63 aminokwasów. We wszystkich badanych przypadkach udało się poprawnie przewidzieć miejsce dokowania ligandów i ich własną konformację w kompleksie (przykład pokazano na rysunku 5). Tak wiec, zaproponowana metoda może być wykorzystywana do przewidywania nieznanych jeszcze struktur kompleksów białko–peptyd i białko–białko.

Dalsze prace nad przedstawionym tu wieloskalowymi sposobami modelowania biomakromolekuł będą zmierzały w kierunku pełnego zautomatyzowania procedur obliczeniowych do przewidywania struktur białek



Rysunek 5. **Kompleks Troponin C — Troponin I. Receptor przedstawiono za pomocą modelu wstęgowego, a ligand — szkieletowego. Model teoretyczny (góra rysunku) i struktura krystalograficzna (dół) są bardzo podobne (1,76 Å RMSD po najlepszym nałożeniu)**

dla całych genomów, opracowania mezoskopowej reprezentacji innych niż białka molekuł oddziałujących z białkami (związki małocząsteczkowe, kwasy nukleinowe, fosfolipidy itd.), a także mezoskopowego modelowania wielkich układów biomakromolekularnych.

# Literatura uzupełniająca


[1] D. Baker, A. Sali, "Protein Structure Prediction and Structural Genomics", *Science*, **294**:93–6 (2001).

[2] A. Kolinski and J. Skolnick, "Reduced Models of Proteins and Their Applications", *Polymer,* **45**:511–524 (2004).

[3] A. Kolinski, "Protein Modeling and Structure Prediction with a Reduced Representation", *Acta Biochimica Polonica*, **51**:349–371 (2004).

[4] A. Kolinski and J.M. Bujnicki, "Gneralized Protein Structure Prediction Based on Combination of Fold-recognition with *de novo* Folding and Evaluation of Models", *Proteins*, **61**(S7):84–90 (2005).

[5] D. Gront and A. Kolinski, "HCPM — Program for Hierarchical Clustering of Protein Models", *Bioinformatics*, **21**:3179–3180 (2005).

[6] D. Gront and A. Kolinski, "BioShell — A Package of Tools for Structural Biology Computations", *Bioinformatics*, **22**:621–622 (2006).

[7] D. Gront, S. Kmiecik and A. Kolinski, "BBQ — Backbone Building from Quadrilaterals. A Fast and Accurate Algorithm for Protein Backbone Reconstruction from Alpha Carbon Coordinates", *J. Comput. Chem.* (w druku).

[8] R. Sicinski, A. Kolinski, P. Rotkiewicz, W. Sicinska, J.M. Prahl, C.M. Smith and H.F. De Luca, "2-Ethyl and 2-Ethylidene Analogs of 1?, 25-Dihydroxy-19--norvitamin $D_3$: Synthesis, Conformational Analysis, Biological Activities, and Docking to the Modeled rVDR Ligand Binding Domain", *J. Medicinal Chem.*, **45**:3366–3380 (2002).

[9] E. Malolepsza, M. Boniecki, A. Kolinski and L. Piela, "Theoretical Model of Prion Propagation: a Misfolded Protein Induces Misfolding", *Proc. Natl. Acad. Sci. USA*, **102**:7835–7840 (2005).

[10] M. Kurcinski and A. Kolinski, "Steps Towards Flexible Docking: Modeling of Three-dimensional Structures of the Nuclear Receptors Bound with Peptide Ligands Mimicking Co-activators' Sequences", *J. Steroid Biochem. and Mol. Biol*. (w druku).


# Abstract


Reduced computer modeling of proteins has now about 30 years history. In spite of enormous increase of computing abilities the reduced models are still very important tools for theoretical studies of proteins. It is shown that the reduced-space modeling can be integrated with a detailed all-atom simulations. Such multiscale approach is crucial for high-resolution protein structure predictions, predictions of protein interactions, computer-aided drug design and study of protein dynamics and thermodynamics.



Andrzej Kolinski, PhD, DSc
Laboratory of Theory of Biopolymers
Faculty of Chemistry, Warsaw University
http://www.biocomp.chem.uw.edu.pl
phone 48-22 822-02-11 ext. 320
fax 48-22 822-59-96

Opracowanie edytorskie: Danuta Czudek-Puchalska, Oficyna Wydawnicza Politechniki Warszawskiej, ul. Polna 50, 00-644 Warszawa, tel. 0-22 234-75-03