Przemysław Porębski

Streszczenie pracy doktorskiej pod tytułem:

„Udoskonalenie procesu budowy modelu oraz udokładniania makrocząsteczkowych struktur krystalicznych”

Struktury krystaliczne pozwalają na szczegółową analizę oddziaływań pomiędzy białkami i ich partnerami oraz między białkami i wiązanymi przez nie małymi cząsteczkami. Z tego względu krystalografia rentgenowska jest ważnym narzędziem służącym do poznawania mechanizmów działania enzymów, powstawania kompleksów białkowych czy do projektowania leków. Dokładność i kompletność modelu jest krytyczna dla dalszej analizy struktury, zwłaszcza w przypadku struktur średnio i nisko rozdzielczych. Kompletność modelu jest szczególnie kluczowaw przypadku analizy wiązania małych cząsteczek, gdzie często niższe obsadzenie małych cząsteczek i ich większy stopień nieuporządkowania przysparzają trudności w interpretacji map gęstości elektronowej. Błędy w procesie interpretacji mapy mogą prowadzić do wniosków niemających funkcjonalnego i fizjologicznego uzasadnienia,a także zwiększają prawdopodobieństwo wystąpienia błędów w badaniach metodami obliczeniowymi (np. dynamiki molekularnej czy badań przesiewowych) i negatywnie wpływają na planowanie dalszych badań i kolejnych eksperymentów krystalograficznych.

Sposobem na usprawnienie procesów interpretacji map i udokładniania modeli struktur krystalicznych, zwłaszcza na początkowych etapach i przy niskorozdzielczych danych, gdzie mapy gęstości elektronowej są niskiej jakości i posiadają niski stosunek szumu do sygnału, jest dalszy rozwój metod budowania modeli. W niniejszej pracy doktorskiej, na podstawie nowo zaprojektowanego algorytmu do modelowania konformacji łańcuchów bocznych, przedstawiono jak rozwój algorytmów wpływa na proces budowania modelu, jakość rozwiązanych struktur i możliwość automatyzacji. Podano też przykłady poprawienia modeli za pomocą przedstawionego algorytmu.

Zaprezentowany w pracy algorytm korzysta z metod używanych dotychczas w teoretycznych procesach przewidywania struktury, uwzględniając dodatkowe informacje dostarczone przez wyniki eksperymentalne. Algorytm podzielono na dwa etapy ze zmiennym problemem kombinatorycznym. W pierwszym etapie wykorzystywana jest bardzo gęsta biblioteka konformerów aminokwasów, którym jest przypisywana energia odpowiadająca dopasowaniu danej konformacji do eksperymentalnej gęstości elektronowej oraz

prawdopodobieństwu zaobserwowania danej konformacji wśród dotychczas rozwiązanych struktur krystalicznych. W tym etapie, w celu przyśpieszenia algorytmu, pominięte są wzajemne oddziaływania pomiędzy sąsiadującymi resztami. Następnie wykorzystywany jest algorytm przeszukiwania grafów A* wraz z teorią DEE ("usuwania ślepych zaułków" z ang. Dead-End Elimination) do osiągnięcia najlepszej kombinacji niekolidujących konformacji. W drugim etapie, tak wybrane konformery służą jako zalążki, wokół których jest rozszerzana przestrzeń poszukiwań. W tym etapie energia danej konformacji jest dodatkowo opisana przez oddziaływania niewiążące - elektrostatyczne i van der Waalsa sparametryzowane w polu siłowym OPLS-UA uwzględniające wzajemne odziaływania pomiędzy sąsiadującymi resztami. Dwuetapowe podejście pozwoliło na osiągnięcie wysokiej dokładności modelowania (>90% poprawnie wymodelowanych konformacji dla wysokorozdzielczych struktur) przy dużej szybkości działania (kilkanaście milisekund dla jednej reszty aminokwasowej)

Ponieważ problem modelowania konformacji łańcuchów bocznych jest zbliżony do problemu projektowania sekwencji białek, opisany algorytm wykorzystano również w procesie rozpoznawania sekwencji białka w trakcie budowania modelu struktury krystalicznej. Oryginalny algorytm został również rozwinięty w celu rozpoznawania i modelowania alternatywnych konformacji łańcuchów bocznych występujących w białkach. Ogólna natura algorytmu pozwala na późniejsze rozszerzenie zastosowań w celu udokładniania struktur kwasów nukleinowych oraz małych cząsteczek.

Algorytm został zaimplementowany w programie *Fitmunk*. *Fitmunk* został przetestowany na wielu nowych strukturach rozwiązywanych w ramach różnych projektów realizowanych przez trzy centra genomiki strukturalnej. Użycie *Fitmunka* bezpośrednio po etapie automatycznego budowania modelu pozwoliło na poprawienie wskaźników udokładnienia R i R_{free} przeciętnie o 1%-2%, a w niektórych przypadkach nawet o 8%.

Przemysław Porębski

Abstract of the thesis:

Improvement of model building and refinement of macromolecular crystal structures

Crystal structures allow for a detailed analysis of protein-protein and protein-small molecule interactions. This makes X-ray crystallography an important tool for the elucidation of mechanisms of enzymatic reactions, studies of protein complexes or drug development. Accuracy and completeness of the model are critical for further structural analysis, especially in the case of medium and low resolution structures. Completeness of the model is a key factor when analyzing the binding of small molecules, which quite often have a lower occupancy and are more disordered. Thus, the corresponding electron density map can be difficult to interpret. Errors in interpretation can result in conclusions that are not physiological or relevant from the functional point of view. These errors also increase the probability of mistakes in computational studies (like molecular dynamics or virtual screening) or in designing further studies or crystallographic experiments.

Further development of a crystallographic model building software is a way to improve a process of map interpretation and model refinement, especially at the early stages of refinement or with low resolution data, where the resulting electron density maps are noisy and of low quality. The model building process, quality of the solved structures and possibility to automate can be improved by the development of a new algorithm for modeling of side chain conformations, which is presented in this thesis. It is also shown how models can be improved by the presented algorithm.

The developed algorithm uses methods from protein structure prediction and incorporates additional experimental data. The algorithm is divided into two steps with different combinatorial problem. During the first step the algorithm uses a very dense library of amino acid conformers, which are scored against an experimental electron density map and a conformer probability derived from existing crystal structures. At this step the pairwise interactions between residues are omitted to speed up the calculations. The A* algorithm and the DEE theory (Dead-End Elimination) are used to achieve an optimal combination of non-colliding side chain conformations. At the next step the conformers selected so far are used as seeds, around which the conformational search space is extended. At this point the conformation energy includes pairwise interactions comprising of the electrostatic and van der

Waals terms from the OPLS-UA force field. The two step approach allowed to achieve high modeling accuracy (>90% of correctly modeled conformations for high resolution structures) together with a high execution speed (several milliseconds per one amino acid residue).

Because the problem of amino acid side chain conformation modeling is similar to a protein sequence design, the described algorithm was adopted for protein sequence recognition during protein model building. The original algorithm was also extended to allow for modeling of alternative conformations of side chains. The general implementation of the algorithm allows for future function extension into refinement of nucleic acids and small molecules.

The algorithm was implemented in a program *Fitmunk*. Fitmunk was tested on many new structures solved under several projects from three structural genomics centers. *Fitmunk* allows for 1%-2% improvement of R and R_{free} on average. For several cases it was possible to achieve 8% improvement.