

JANUSZ M. BUJNICKI

*Pracownia Bioinformatyki i Inżynierii Białka
Międzynarodowy Instytut Biologii Molekularnej i Komórkowej (IIMCB)
Trojdena 4, 02-109 Warszawa
Pracownia Bioinformatyki
Instytut Biologii Molekularnej i Biotechnologii
Uniwersytet Adama Mickiewicza
Umultowska 89, 61-614 Poznań
e-mail: iamb@genesilico.pl*

PRZEWIDYWANIE STRUKTURY BIAŁEK: PODEJŚCIE BOLTZMANNOWSKIE I DARWINOWSKIE

WSTĘP

Znajomość struktury białka jest niezbędna do pełnego zrozumienia jego funkcji biochemicznej i komórkowej. Struktury białek o wysokiej rozdzielczości otrzymać można przy pomocy metod doświadczalnych: krystalografii rentgenowskiej i jądrowego rezonansu magnetycznego (NMR). Metody te są jednak czasochłonne i kosztowne oraz wymagają otrzymania kryształów o wysokiej jakości lub stężonego roztworu białka, co dla wielu białek jest trudne lub wręcz niemożliwe. Z tego względu, struktury udało się rozwiązać, jak dotąd, dla mniej niż 1% białek o znanych sekwencjach. Dla przykładu – w maju 2005 r. znanych było około 2500000 sekwencji i jedynie około 22000 struktur o unikalnych sekwencjach [aktualne wartości można sprawdzić w bazach danych NCBI (BENSON i współaut. 2005) i Protein Data Bank (BERMAN i współaut. 2000)]. Jeżeli weźmie się do tego pod uwagę, że rozmiar bazy danych sekwencji podwaja się mniej więcej co 18 miesięcy, podczas gdy ilość znanych struktur podwaja się mniej więcej co 3 lata, to staje się jasne, że w przewidywalnej przyszłości nie uda się rozwiązać struktur wszystkich białek, a olbrzymi rozróżnienie pomiędzy liczbą znanych sekwencji i rozwiązanych struktur będzie się nadal powiększał.

Ponad 40 lat temu Anfinsen wykazał, że cała informacja niezbędna do przyjęcia przez białko jego struktury natywnej dla danych

warunków jest zakodowana w jego sekwencji aminokwasowej (ANFISEN i współaut. 1961). To odkrycie, uhonorowane nagrodą Nobla, stanowi podstawę hipotezy, że struktura białka odpowiada globalnemu minimum energii swobodnej i może być wyliczona (wymodelowana) w oparciu o znajomość sekwencji i matematyczno-fizyczny model oddziaływań pomiędzy atomami cząsteczki białka i pomiędzy białkiem i otoczeniem. Rozwijająca się obecnie inicjatywa „genomiki strukturalnej” stawia sobie za cel doświadczalne rozwiązanie struktury jedynie dla najważniejszych bądź dla najbardziej reprezentatywnych białek. Dla pozostałych białek, czyli dla olbrzymiej większości, proponuje się zastosowanie metod modelowania teoretycznego (BAKER i SALI 2001). Mimo iż zadanie wiarygodnego i dokładnego teoretycznego przewidywania struktury białka na podstawie znajomości jego sekwencji pozostaje nieosiągniętym „Świętym Graalem” współczesnej nauki, ostatnio poczyniono duże postępy, które umożliwiają budowę użytecznych modeli teoretycznych przynajmniej dla części białek.

DWIE SZKOŁY MODELOWANIA STRUKTURY BIAŁKA

Można wyróżnić dwa podstawowe podejścia do przewidywania struktury białka, które polegają na komputerowym modelowaniu elementów procesów o zupełnie różnej skali

czasowo-przestrzennej: zwijania się łańcucha polipeptydowego lub ewolucji białka. Badacze zajmujący się problematyką przewidywania struktury białka zwykle identyfikują się z jedną z dwóch dyscyplin: fizyką statystyczną, przy pomocy której można opisywać proces zmian konformacji i energii swobodnej

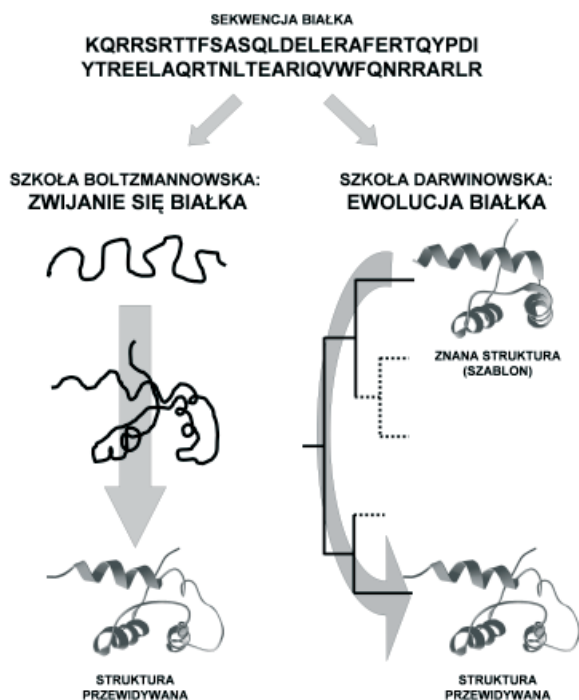
białka w skali ułamków sekundy (nazywana jest ona czasem „szkołą Boltzmannowską”) lub z biologią ewolucyjną, przy pomocy której można opisywać proces zmian sekwencji i struktury białka zachodzący w przyrodzie w skali milionów lat (nazywana jest ona czasem „szkołą Darwinowską”) (Ryc. 1).

PODEJŚCIE FICZYCZNE – SZKOŁA BOLTZMANNOWSKA

Twierdzenie Anfinsena, że struktura większości białek jest zdeterminowana wyłącznie przez ich sekwencję aminokwasową i odpowiada minimum energii swobodnej układu, w którym znajduje się białko, jest powszechnie akceptowane (choć znane są wyjątki – patrz BAKER i AGARD 1994). Zatem znajomość sekwencji aminokwasowej białka powinna wystarczyć do obliczenia konformacji o najniższej energii. Teoretycznie wystarczy komputerowo wygenerować wszystkie możliwe konformacje łańcucha polipeptydowego,

następnie obliczyć energię swobodną tych konformacji w układzie odpowiadającym środowisku komórki i zidentyfikować strukturę o najniższej energii. Liczba możliwych konformacji nawet dla białka o krótkim łańcuchu jest jednak astronomicznie wielka. Już w latach 70. XX w. Levinthal zauważył, że modelowe białko o 150 aminokwasach posiada 450 stopni swobody i gdyby chciał oszacować energię dla konformacji różniących się o $1/10$ radiana, to należałoby przeanalizować ok. 10^{300} struktur (KARPLUS 1997). W komórce proces zwijania się białka, czyli przejście od konformacji przypadkowej do ściśle określonej formy natywnej, wymaga czasu rzędu milisekund do sekund i oczywistym jest, że białko nie „testuje” wszystkich możliwych konformacji w poszukiwaniu globalnego minimum energetycznego, bo proces ten mógłby okazać się dłuższy niż wiek wszechświata. Z podobnych względów niemożliwe jest więc wygenerowanie wszystkich konformatorów i obliczenie ich energii przy pomocy nawet najszybszych superkomputerów świata.

Nawet jeżeli nie analizuje się wszystkich konformacji, a jedynie symuluje dynamikę molekularną układu, w której *explicite* analizuje się oddziaływania wszystkich atomów białka i roztworu, to komputerowe symulacje, które odpowiadają czasowi rzeczywistemu rzędu mikrosekund, zajmują w najlepszym razie wiele miesięcy. Jest to wciąż co najmniej tysiąc razy za wolno. Aby przyspieszyć obliczenia i umożliwić symulacje całego procesu zwijania się białka, podejmuje się próby zastosowania uproszczonych modeli białek, w których całe grupy atomów traktuje się jako zunifikowane „pseudoatomy”. Upraszcza to opis geometrii molekuł białka i zarazem przyspiesza obliczanie energii modelowych struktur. Innym sposobem na przyspieszenie obliczeń jest dyskretyzacja przestrzeni konformacyjnej (np. poprzez umieszczanie atomów lub pseudoatomów) na węzłach trójwymiarowej siatki. Dzięki temu liczba możliwych konformacji, choć wciąż



Ryc. 1. "Fizyczne" i „ewolucyjne” podejście do przewidywania struktury białka.

Dla zadanej sekwencji aminokwasowej, można przeprowadzić symulację procesu zwijania się łańcucha polipeptydowego z użyciem modelu oddziaływań fizykochemicznych (szkoła „Boltzmannowska”) lub symulację procesu ewolucji w oparciu o znaną strukturę spokrewnionego białka, z użyciem modelu dywergentnych zmian sekwencji i struktury (szkoła „Darwinowska”).

ogromna – staje się skończona. Skończona staje się też liczba możliwych lokalnych konfiguracji (wzajemnych orientacji atomów i pseudoatomów różnego typu), można zatem obliczyć energię wszystkich możliwych typów oddziaływań lokalnych i podczas analizy poszczególnych konformacji tylko wczytywać ją, zamiast obliczać za każdym razem od nowa (KOLINSKI 2004).

Jednym z najczęściej używanych podejść do przeszukiwania przestrzeni konformacyjnej w celu znalezienia globalnego minimum energetycznego (zarówno w modelach ciągłych, jak i siatkowych) jest tzw. dynamika Monte Carlo (MC), polegająca na wprowadzaniu losowych zmian we fragmentach struktury i obliczaniu energii nowej konformacji. Kroki prowadzące do konformacji o niższej energii są zawsze akceptowane, natomiast kroki prowadzące do wyższych temperatur są akceptowane lub odrzucane z prawdopodobieństwem wynikającym z rozkładu Boltzmann (stąd właśnie nazwa „szkoły”), a zależnym od różnicy energii pomiędzy obydwo ma konformacjami i od temperatury, w jakiej prowadzona jest symulacja. W ten sposób łańcuch białka jest w stanie pokonywać bariery pomiędzy lokalnymi minimami energetycznymi (z prawdopodobieństwem proporcjonalnym do wysokości bariery), a także badać jak głębokie jest każde minimum.

Do identyfikacji struktury natywnej (lub bliskiej do natywnej) niezbędna jest dokładna i wiarygodna funkcja oceniająca energię swobodną układu. Niestety, nasza umiejętność obliczania energii oddziaływań wewnątrz cząsteczki białka i pomiędzy białkiem i jego otoczeniem jest wciąż niedoskonała – jak dotąd nie udało się opracować metody *ab initio*, tzn. opierającej się wyłącznie na prawach fizyki, która poprawnie, dokładnie i wiarygodnie przewidywałaby natywną konformację białka. Stosując to podejście, w najlepszym razie udaje się przewidzieć strukturę peptydów o długości najwyżej kilkunastu aminokwasów, a i tak rzadko można mieć pewność, że struktura wymodelowana *ab*

initio jest poprawna. Ze względu na problemy ze zdefiniowaniem wystarczająco dokładnej funkcji energii opisującej oddziaływania wewnątrz białka i białka ze środowiskiem, w wielu metodach symulujących zwijanie się białka wprowadzono alternatywne rozwiązanie. Zamiast energii „fizycznej” oblicza się „pseudoenergię”, czyli potencjał statystyczny wyprowadzony z analizy częstości występowania oddziaływań danego typu w uprzednio poznanych strukturach. Wykazano, że w wielu zastosowaniach potencjały statystyczne prowadzą do jakościowo i ilościowo podobnych wyników, co „ortodoksyjne” potencjały fizyczne. Co więcej, niektóre typy oddziaływań, których mechanizm nie jest wystarczająco dobrze zrozumiany (np. efekt hydrofobowy), są zwykle dokładniej modelowane przez potencjały statystyczne (LAZARIDIS i KARPLUS 2000).

Niestety, nawet najlepsze spośród istniejących metod przeszukiwania przestrzeni konformacyjnej i potencjałów dowolnego typu są, ogólnie rzecz biorąc, niezdolne do generowania wiarygodnych modeli, podobnych do struktury natywnej białka. Nawet jeżeli przyjmemy, że istniejące funkcje energii są w stanie w wielu wypadkach odróżnić konformację natywną od konformacji nienatwinych, to podczas symulacji zwykle nie udaje się wygenerować konformacji na tyle bliskich konformacji natywnej, że zostałyby one zidentyfikowane w oparciu o kryterium energetyczne. Prowadząc odpowiednio długie symulacje udaje się zwykle wygenerować zbiór konformacji (rzędu np. dziesiątek tysięcy lub więcej alternatywnych struktur), wśród których jedna lub więcej przypomina konformację natywną. Jednak nawet dla bardzo dobrych modeli, których średnie odchylenie kwadratowe koordynat od struktury natywnej (RMSD) jest mniejsze od 3 Å, „z punktu widzenia potencjału energetycznego” różnica ta jest na tyle duża, że inne konformacje (nawet zupełnie nie przypominające konformacji natywnej) mogą wykazywać porównywalną lub niższą energię.

PODEJŚCIE EWOLUCYJNE – SZKOŁA DARWINOWSKA

Badania nad strukturą białek spokrewnionych ewolucyjnie (homologicznych) wykazały, że w procesie ewolucji dywergentnej polegającej na nagromadzaniu mutacji (substytucji, insercji, delecji) podobieństwo struktury maleje znacznie wolniej niż podobieństwo

sekwencji (CHOTHIA i LESK 1986). Znanych jest wiele białek wzajemnie homologicznych, które zachowały uderzające podobieństwo strukturalne pomimo niemal całkowitej utraty podobieństwa sekwencji. Zdarza się, że w niektórych rodzinach białek zachowywa-

na jest struktura i tylko kilka aminokwasów niezbędnych dla funkcji białka, np. biorących udział w przeprowadzaniu reakcji enzymatycznej. Ogólnie przyjmuje się regułę, że białka homologiczne zachowują podobieństwo struktury w obrębie poszczególnych domen strukturalnych; większość zmian konformacyjnych obserwowanych pomiędzy homologami dotyczy pętli i innych elementów na powierzchni białka, podczas gdy rdzeń zmienia się w bardzo niewielkim stopniu.

Powyższą regułę empiryczną wykorzystano do opracowania podejścia zwanego „modelowaniem homologicznym”, w którym strukturę białka przewiduje się nie poprzez symulację zwijania łańcucha (jak w podejściu fizycznym, „Boltzmannowskim”), ale poprzez symulację procesu ewolucji sekwencji i struktury. Tak jak podejście „Boltzmannowskie” opiera się na modelu matematyczno-fizycznym, opisującym zależności między konformacją łańcucha polipeptydowego i energią, podejście „Darwinowskie” opiera się na modelu opisującym zależności pomiędzy zmianami w sekwencji aminokwasowej i ich wpływem na zmianę lokalnej konformacji białka. Kluczowa różnica pomiędzy obydwo-ma metodami polega na tym, że „metoda Boltzmannowska” usiłuje przewidzieć strukturę natywną opierając się wyłącznie na sekwencji badanego białka, podczas gdy „metoda Darwinowska” wymaga dodatkowo znajomości struktury innego, spokrewnionego białka, które może posłużyć jako „szablon” do modelowania struktury badanego białka o nieznannej strukturze – „celu” (BYSTROFF i SHAO 2004).

Proces modelowania struktury białka „metodą Darwinowską” można z grubsza podzielić na trzy etapy: (1) identyfikacja spokrewnionego białka o znanej strukturze (szablonu), (2) modelowanie procesu ewolucji sekwencji poprzez przyrównanie (*ang. alignment*) sekwencji celu i szablonu tak, aby aminokwasy pochodzące od wspólnego przodka znalazły się w parach, a aminokwasy wstawione i usunięte w procesach insercji i delekcji nie miały pary, (3) wprowadzenie substytucji, insercji i delekcji aminokwasów oraz optymalizacja tak powstałego modelu, żeby „nowe” aminokwasy nie nachodziły na siebie i aby oddziaływania między nimi przypominały oddziaływania znane z innych białek. Etapy (1) i (2) bardzo często łączą się w jeden proces polegający na przeszukaniu bazy danych sekwencji (i/lub struktur) i wygenerowanie przyrównań cel-szablon. Obec-

nie rezygnuje się najczęściej z porównywania sekwencji parami, ponieważ jest ono bardzo mało czułe i umożliwia identyfikację tylko bardzo bliskich homologów. Do identyfikacji szablonów i generowania przyrównań używa się metod rozpoznających zwój (*ang. fold-recognition, FR*), które dzielą się na dwie klasy: wykorzystujące informacje o strukturze trzeciorzędowej szablonu (tzw. metody „przewlekania”, *ang. threading*) lub pomijające ją i badające wyłącznie podobieństwo sekwencyjne (CYMERMAN i współaut. 2004).

Pierwsze metody przewlekania używały potencjałów energetycznych (fizycznych lub statystycznych, podobnych do tych używanych przez „boltzmannowskie” metody do zwijania struktury białka *de novo*), aby obliczyć energię oddziaływania aminokwasów celu, gdy badana sekwencja dopasowana jest optymalnie do „rusztowania”, jakie stanowi struktura szablonu. Praktyka wykazała jednak, że takie metody są mało skuteczne – tak jak w przypadku metod symulujących zwijanie białka, ocena energii oddziaływań jest po prostu zbyt niedoskonała. Najlepsze istniejące obecnie metody wykorzystują informacje o konserwacji sekwencji zarówno celu jak i każdego potencjalnego szablonu. Z reguły wyszukuje się sekwencje białek blisko spokrewnionych z celem, buduje się dla nich przyrównanie wielosekwencyjne (*ang. multiple alignment*), a następnie oblicza macierz opisującą częstość występowania 20 aminokwasów w poszczególnych pozycjach lub ukryty model Markowa (*ang. Hidden Markov Model, HMM*), który jest innym sposobem na probabilistyczne zakodowanie tej samej informacji. Następnie profil lub HMM przyrównuje się do profili lub HMM zbudowanych w podobny sposób dla wszystkich białek o znanych strukturach i identyfikuje te z nich, które wykazują największe podobieństwo do celu.

Metody służące rozpoznawaniu zwinięcia białka są w stanie poprawnie zidentyfikować odpowiedni szablon do modelowania w około połowie przypadków, gdy sekwencja celu nie wykazuje statystycznie istotnego podobieństwa do żadnego białka o znanej strukturze. Słabym punktem wszystkich tych metod (zarówno opartych o same sekwencje jak i o sekwencje i struktury) jest brak skutecznej funkcji oceniającej przyrównania sekwencji celu i szablonu. Zdarza się, że potencjalny szablon o strukturze podobnej do prawdziwej struktury białka-celu istnieje, ale algorytm nie potrafi poprawnie przyrównać

sekwencji celu do tego szablonu albo ocenia to przyrównanie gorzej niż dopasowanie do innych, niepodobnych struktur (błędnych szablonów). W sytuacji, kiedy nie jest się pewnym szablonu lub przyrównania, najlepszym rozwiązaniem jest wygenerowanie modelu pełnoatomowego i ocena jego struktury na poziomie atomowym.

Wygenerowanie pełnoatomowej reprezentacji celu polega zazwyczaj na skopiowaniu tych elementów szablonu, które są konserwowane w celu oraz dodaniu elementów, które w celu są inne niż w szablonie. Konserwowane elementy mogą być kopiowane *explicite* – wtedy produktem pośrednim jest niepełny model zbudowany z części wspólnej szablonu i celu. Uzupełniany jest on następnie przez wstawienie segmentów łańcucha odpowiadających insercjom i połączeniu końców powstałych w miejscach delekcji (konformacje tych „nowych” segmentów mogą być pobrane ze znanych struktur lub wygenerowane „*de novo*”) oraz ustaleniu rotamerów łańcuchów bocznych, które uległy substytucji – w ten sposób działa np. SWISS-MODEL (SCHWEDE i współaut. 2003). Innym podejściem jest wyprowadzenie więzów określających odległości pomiędzy atomami wspólnymi dla celu i szablonu, a następnie zwinięcie struktury celu tak, aby zminimalizować naruszenie tych więzów a także ograniczeń dotyczących akceptowalnych wartości kątów i długości wiązań, ustalonych poprzez analizę konformacji w znanych strukturach – w ten sposób działa MODELLER (SALI i BLUNDELL 1993).

Metody ewolucyjne są w stanie wygenerować modele o wysokiej jakości (porównywalnej z jakością struktur NMR) jeżeli podobieństwo sekwencji między celem i szablonem jest bardzo wysokie (identyczność aminokwasów > 50%). Przy obniżającym się podobieństwie celu i szablonu liczba błędów w przyrównaniu ich sekwencji rośnie lawinowo, co znacząco obniża dokładność modelu. Warto przy tym dodać, że metody „czysto ewolucyjne” z reguły nie są w stanie wygenerować modelu, w którym konformacja łańcucha głównego byłaby bardziej podobna do prawdziwej struktury celu, niż konformacja szablonu. To ograniczenie można ominąć (przynajmniej w teorii) poprzez połączenie podejścia ewolucyjnego i fizycznego. Zazwyczaj polega to na wykorzystaniu metod ewolucyjnych do identyfikacji najbardziej prawdopodobnego zwoju, w ramach którego poszukuje się minimum energetycznego

metodami zawierającymi przynajmniej elementy podejścia fizycznego. Alternatywnym podejściem jest przeszukiwanie przestrzeni konformacyjnej poprzez składanie modelu celu z fragmentów znanych struktur, które poniekąd nawiązuje do hipotezy, iż pierwotne białka wyewoluowały przez asocjację krótkich peptydów (LUPAS i współaut. 2001, BUJNICKI 2005).

Bardzo ważnym etapem, choć niestety często zaniechanym, jest ocena poprawności modelu. Dostępnych jest wiele metod służących ocenie poprawności struktury białka, które zostały stworzone na potrzebę modeli krystalograficznych. Te z nich, które oceniają na ile model jest „białko-podobny”, tzn. na ile jego cechy odpowiadają wartościom typowym dla znanych (i niewątpliwie poprawnych) struktur, mogą być również stosowane do analizy modeli teoretycznych. Metody te najczęściej analizują cechy strukturalne, takie jak geometria, stereochemia, niektóre z nich używają również potencjałów statystycznych oceniających zgodność charakteru fizykochemicznego poszczególnych aminokwasu z otoczeniem w jakim zostały umieszczone w modelu. Należy podkreślić, że ocena poprawności modelu wyłącznie pod względem parametrów stereochemicznych (np. popularna w badaniach krystalograficznych analiza wykresu Ramachandrana) ma raczej niewielki sens dla modeli homologicznych, bo parametry te są najczęściej skopiowane z użytych szablonów strukturalnych i odzwierciedlają w większym stopniu jakość szablonu niż modelu. Na przykład można łatwo wygenerować zupełnie błędny model struktury białka wykazujący doskonałą stereochemię (np. przez błąd w przyrównaniu sekwencji celu do szablonu), jak i model bliski strukturze natywnej, w którym długości i kąty wiązań będą jednak dalekie od idealnych. Znacznie trudniej „oszukać” jest potencjały statystyczne, podobne do tych używanych przez metody używane do zwijania białek, które w tym przypadku kładą nacisk na ocenę lokalnego środowiska (np. czy hydrofobowe grupy boczne otoczone są innymi grupami hydrofobowymi i czy naładowane grupy boczne są eksponowane na zewnątrz białka lub sparowane z grupami bocznymi o przeciwnym ładunku). Należy podkreślić, że wymienione metody oceniają jedynie na ile dany model jest podobny do wcześniej znanych struktur, ale nie są w stanie obiektywnie stwierdzić, czy jest on poprawny, czy błędny. W szczególności, za „błędne” mogą

zostać uznane miejsca ważne dla funkcji białka, np. eksponowane na powierzchni aminokwasy hydrofobowe biorące udział w wiązaniu innego białka, lub skupione w jednym miejscu aminokwasy naładowane ujemnie, które w warunkach fizjologicznych są zobojętnione przez wiązanie jonów metalu itp. Problem ten zminimalizować można poprzez weryfikację modelu z użyciem kilku różnych metod, które kładą nacisk na różne parametry strukturalne (SASIN i BUJNICKI 2004).

Najlepszym rozwiązaniem pozostaje jednak weryfikacja modelu strukturalnego z użyciem danych doświadczalnych, najlepiej takich, z których można wyprowadzić więzy odległościowe (np. modyfikacja chemiczna, sieciowanie, ewentualnie mutageneza). Aby weryfikacja miała znaczenie statystyczne, niezbędne jest wygenerowanie alternatywnych modeli i ocena na ile dany zbiór więzów faworyzuje dany model względem alternatywnych konformacji (YE i współaut. 2004).

MARIAŻ BOLTZMANNA I DARWINA

Najlepsze modele struktury białka używa się obecnie metodą „meta”, czyli przez użycie wielu różnych metod i zastosowanie zasady konsensusu (BUJNICKI i FISCHER 2004). W praktyce polega to na użyciu możliwie dużej liczby różnych metod FR do wygenerowania alternatywnych przyrównań pomiędzy sekwencją celu i białkami o znanych strukturach, wśród których identyfikuje się najczęściej występujący zwój i jeden lub więcej najbardziej obiecujących szablonów. W oparciu o te przyrównania buduje się serię alternatywnych modeli, których jakość należy koniecznie oszacować, najlepiej przy użyciu kilku różnych metod, z naciskiem najpierw na globalną poprawność modelu, a następnie na „białkopodobność” lokalnych konformacji. Jeżeli udaje się zidentyfikować w miarę dobrze ocenione modele o globalnie podobnych strukturach, to można założyć, że z dużym prawdopodobieństwem udało się odgadnąć zwój badanego białka. Na tym etapie model można poprawić na kilka sposobów, m.in. poprzez lokalne modyfikacje przyrównania w regionach odpowiadającym relatywnie nisko ocenionym konformacjom lub poprzez zastosowanie innej procedury modelowania. Niekompletny model „Darwinowski”, np. taki, w którym brakuje końców lub długich pętli na skutek braku odpowiednich struktur w szablonach, można uzupełnić przy użyciu metod „Boltzmannowskich”, które przeprowadzają lokalne przeszukiwanie przestrzeni konformacyjnej dla brakujących elementów i identyfikują takie warianty, które wykazują optymalną energię oddziaływań z pozostałą częścią modelu.

Wiele struktur zbudowanych metodami „Darwinowskimi” jest wystarczająco poprawnych i dokładnych, aby stanowić dogodny punkt wyjścia do rozważań dotyczących

funkcji badanego białka i może pomóc np. w identyfikacji aminokwasów odpowiedzialnych za stabilność lub oddziaływanie z innymi cząsteczkami w komórce. Należy jednak pamiętać, że modele optymalne w sensie ewolucyjnym (zbudowane w oparciu o idealne przyrównanie do szablonu) właściwie nigdy nie są optymalne w sensie fizycznym, tzn. nie znajdują się w globalnym minimum energetycznym. Dlatego standardowa minimalizacja energii całego modelu „Darwinowskiego” praktycznie zawsze prowadzi do pogorszenia jego jakości, na skutek wprowadzenia go w lokalne, a nie globalne minimum energetyczne. Tym niemniej, dobre modele „Darwinowskie” mają konformację stosunkowo bliską konformacji natywnej i z tego powodu mogą służyć jako punkt wyjścia do lokalnego przeszukiwania przestrzeni konformacyjnej metodami „Boltzmannowskimi”. Niestety, nawet po ograniczeniu przestrzeni konformacyjnej do bliskiego sąsiedztwa modelu Darwinowskiego, liczba możliwych struktur pozostaje zbyt wielka, by można było obliczyć energię dla wszystkich. Z tego względu ostatnio intensywnie rozwija się metody przewidywania możliwych globalnych zmian konformacyjnych w oparciu o analizę plastyczności białka — zarówno z perspektywy ewolucji (zmiennosc w rodzinie spokrewnionych białek), jak i z perspektywy fizyki (naturalne fluktuacje struktury jak i zmiany zachodzące pod wpływem wiązania innych cząsteczek) (LEO-MACIAS i współaut. 2005). Postęp osiągnięty ostatnio w tej dziedzinie (MISURA i BAKER 2005) daje nadzieję, że wkrótce będzie możliwe zwiększenie dokładności zgrubnie poprawnych modeli „Darwinowskich” do poziomu zbliżonego do modeli średniej rozdzielczości rozwiązanych metodami krystalograficznymi lub przez NMR.

Jeżeli uda się osiągnąć sukces w, wydawałoby się, tak łatwym zadaniu jak poprawianie już całkiem dobrych modeli „Darwinowskich”, to jest nadzieja, że opracowane w tym celu metody do przeszukiwania przestrzeni konformacyjnej oraz funkcje energii będą się również nadawały do wiarygodnego przewidywania struktury białka „*ab initio*” wyłącznie w oparciu o prawa fizyki. Stanowiłoby to od dawna oczekiwany przełom w badaniach nad mechanizmami zwijania się białka

i zapewne umożliwiłoby w pełni racjonalną inżynierię białek o zupełnie nowych strukturach i funkcjach.

Autor chciałby podziękować członkom swojej grupy badawczej za pomocne uwagi podczas pisanego tego artykułu oraz wszystkim współpracownikom za współudział w badaniach nad rozwojem i zastosowaniami metod przewidywania struktury białek.

PROTEIN STRUCTURE PREDICTION: BOLTZMANNIAN AND DARWINIAN APPROACHES

Summary

Efforts to solve the problem of protein folding have been traditionally rooted in two schools of thought named „Boltzmannian” and „Darwinian”, after the scientists who defined the fundamental principles of statistical thermodynamics and evolutionary biology, respectively. One approach to protein structure prediction is based on the principles of physics, e.g. on the thermodynamic hypothesis, according to which the native structure of a protein corresponds to the global minimum of its free energy under given conditions. Accordingly, the physics-based methods model the process of protein folding by simulating the conformational changes and searching for the free energy minimum. The other approach is

based on the principles of evolution, in particular the empirical rule that evolutionarily related (homologous) proteins usually retain the same three-dimensional fold despite the accumulation of divergent mutations. Evolution-based methods attempt to map the sequence of the target protein to the structure of another protein (a template), model the overall fold of the target based on that of the template and infer how the target structures will change due to substitutions, insertions and deletions, as compared with the template. This review summarizes the basics of protein structure prediction by both types of approaches and discusses the issue of model quality evaluation.

LITERATURA

- ANFINSEN C. B., HABER E., SELA M., WHITE F. H., JR., 1961. *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain*. Proc. Natl. Acad. Sci. USA 47, 1309–1314.
- BAKER D., AGARD D. A., 1994. *Kinetics versus thermodynamics in protein folding*. Biochemistry 33, 7505–7509.
- BAKER D., SALI A., 2001. *Protein structure prediction and structural genomics*. Science 294, 93–96.
- BENSON D. A., KARSCH-MIZRACHI I., LIPMAN D. J., OSTELL J., WHEELER D. L., 2005. *GenBank*. Nucleic Acids Res. 33, D34–38.
- BERMAN H. M., WESTBROOK J., FENG Z., GILLILAND G., BHAT T. N., WEISSIG H., SHINDYALOV I. N., BOURNE P. E., 2000. *The Protein Data Bank*. Nucleic Acids Res. 28, 235–242.
- BUJNICKI J. M. 2005. *Protein structure prediction by recombination of fragments*. ChemBioChem (w druku).
- BUJNICKI J. M., FISCHER D., 2004. *‘Meta’ approaches to protein structure prediction*. [W:] *Practical Bioinformatics*. BUJNICKI J. M. (red.). Springer-Verlag, Berlin, 23–34.
- BYSTROFF C., SHAO Y., 2004. *Modeling protein folding pathways*. [W:] *Practical Bioinformatics*. BUJNICKI J. M. (red.). Springer-Verlag, Berlin, 97–122.
- CHOTHIA C., LESK A. M., 1986. *The relation between the divergence of sequence and structure in proteins*. EMBO J. 5, 823–826.
- CYMERMAN I. A., FEDER M., PAWŁOWSKI M., KUROWSKI M. A., BUJNICKI J. M., 2004. *Computational methods for protein structure prediction and fold recognition*. [W:] *Practical Bioinformatics*. BUJNICKI J. M. (red.). Springer-Verlag, Berlin, 1–21.
- KARPLUS M., 1997. *The Levinthal paradox: yesterday and today*. Fold. Des. 2, S69–S75.
- KOLINSKI A., 2004. *Protein modeling and structure prediction with a reduced representation*. Acta Biochim. Polon. 51, 349–371.
- LAZARIDIS T., KARPLUS M., 2000. *Effective energy functions for protein structure prediction*. Curr. Opin. Struct. Biol. 10, 139–145.
- LEO-MACIAS A., LOPEZ-ROMERO P., LUPYAN D., ZERBINO D., ORTIZ A. R., 2005. *Core deformations in protein families: a physical perspective*. Biophys. Chem. 115, 125–128.
- LUPAS A. N., PONTING C. P., RUSSELL R. B., 2001. *On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?* J. Struct. Biol. 134, 191–203.
- MISURA K. M., BAKER D., 2005. *Progress and challenges in high-resolution refinement of protein structure models*. Proteins 59, 15–29.
- SALI A., BLUNDELL T. L., 1993. *Comparative protein modelling by satisfaction of spatial restraints*. J. Mol. Biol. 234, 779–815.
- SASIN J. M., BUJNICKI J. M., 2004. *COLORADO3D, a web server for the visual analysis of protein structures*. Nucleic Acids Res. 32, W586–W589.
- SCHWEDE T., KOPP J., GUERX N., PEITSCH M. C., 2003. *SWISS-MODEL: An automated protein homology-modeling server*. Nucleic Acids Res. 31, 3381–3385.

YE X., O'NEIL P. K., FOSTER A. N., GAJDA M. J., KOSINSKI J., KUROWSKI M. A., BUJNICKI J. M., FRIEDMAN A. M., BAILEY-KELLOGG C., 2004. *Probabilistic cross-*

link analysis and experiment planning for high-throughput elucidation of protein structure. Protein Sci. 13, 3298–3313.