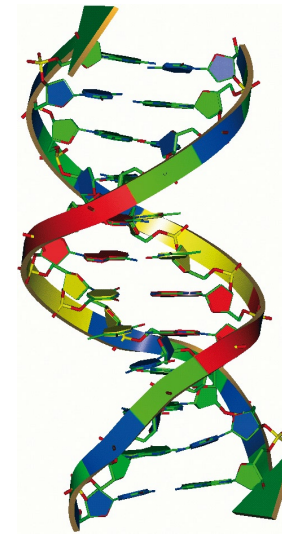
prof. dr hab. inż. Jacek Błażewicz
Instytut Informatyki, Politechnika Poznańska

BIOINFORMATYKA I JEJ PERSPEKTYWY

1. Wstęp

Bioinformatyka jest jedną z najmłodszych nauk, której burzliwy rozwój został wymuszony przez postęp w dziedzinie nauk biologicznych, a umożliwiły go dokonane przełomowe osiągnięcia i wdrożenia w samej informatyce. Tak jak wiek XIX był wiekiem węgla, pary i mechaniki, a wiek XX – wiekiem atomu i fizyki, tak wszystko wskazuje na to, że wiek XXI będzie wiekiem *biologii* i właśnie *informatyki*. Świadczy o tym między innymi fakt, iż większość przełomowych prac publikowanych w ostatnich dziesięcioleciach w prestiżowych czasopismach naukowych *Science* i *Nature* dotyczy biologii, a do uzyskania tych rezultatów w większości przypadków wykorzystano skomplikowane narzędzia bioinformatyczne.

Przełomowym odkryciem II połowy XX wieku, stymulującym powyższy rozwój nauki było podanie w 1953 r. przez Watsona i Cricka (wykorzystujących także prace Wilkinsa i Franklin) *modelu podwójnej helisy łańcucha DNA* (por. Rys. 1), przechowującego (kodującego) informację genetyczną we wszystkich organizmach żywych [41], [19] i [42]. Za osiągnięcie to badacze uzyskali kilka lat później nagrodę Nobla. Odkrycie budowy helisy zrewolucjonizowało nasze spojrzenie na biologię, w szczególności biologię molekularną, której rozwój znacząco przyspieszył. Jej wpływ na informatykę też był bardzo znaczący. Po pierwsze, ilość generowanych danych z eksperymentów biologicznych okazała się za duża do zapamiętania i przetworzenia „ręcznego". Niezbędna była pomoc informatyki. Z drugiej strony łańcuch DNA okazał się *zakodowaną informacją* dotyczącą budowy i funkcjonowania organizmów



Rys. 1. Podwójna helisa DNA.

żywych. Niezbędne okazało się zatem wykorzystanie informatyki co najmniej na dwóch poziomach pracy biologów:

- na etapie *gromadzenia* i *efektywnego zapisywania, przechowywania i wydobywania informacji*, co prowadziło do wykorzystywania istniejących, a później stymulacji rozwoju nowych systemów baz danych;
- na etapie *przetwarzania informacji* zawartej w łańcuchach DNA, a także RNA i białek, modelowania struktur przestrzennych tych molekuł, a także przewidywania ich funkcjonalności i wzajemnych powiązań, co wymagało zastosowania dostępnych i rozwoju nowych narzędzi informatycznych i matematycznych, w szczególności, kombinatoryki, probabilistyki, statystyki, geometrii przestrzennej i wielu innych.

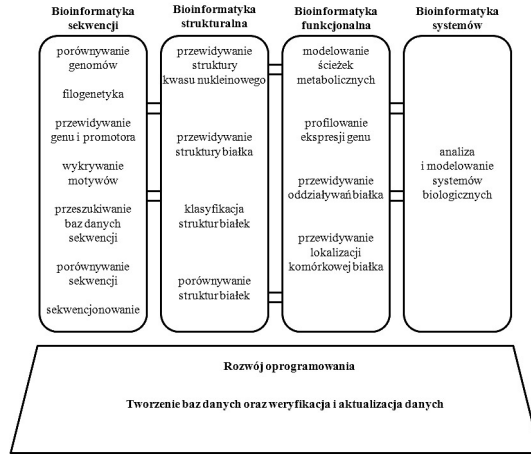
Możemy zatem powiedzieć, że *bioinformatyka* to dziedzina wiedzy, która sytuuje się na pograniczu informatyki, biologii i matematyki. Wymaga znajomości tych trzech obszarów badawczych, ale w tej chwili jest traktowana jako samodzielna dyscyplina naukowa. Przepływ idei jest tu następujący: biologia, która jak powiedzieliśmy, jest w tej chwili motorem napędowym nauki, kreuje pewien problem, a informatyka próbuje (wykorzystując także matematykę) pomóc w jego rozwiązaniu. Należy podkreślić tutaj istotność aspektu obliczeniowego, a więc wykorzystanie komputerów. Samo zamodelowanie pewnego zjawiska, np. przy wykorzystaniu równań matematycznych, nie wystarczy! Potrzebne są efektywne algorytmy i programy komputerowe konstruujące rozwiązanie, choćby nie zawsze optymalne. To między innymi przyczyna wielkiego sukcesu bioinformatyki i stosunkowo niewielkiego zainteresowania biomatematyką. Warto w tym miejscu również wspomnieć o niesłychanym postępie nauk obliczeniowych (nie tylko związanych z biologią i bioinformatyką), który został spowodowany rozwojem Internetu i wielu różnorodnych technik obliczeń równoległych.

Wspomniany wyżej aspekt konstrukcyjny doprowadził też do często występującego podziału interesującej nas dziedziny na *biologię obliczeniową* i właściwą *bioinformatykę*, w zgodzie z obowiązującym rozróżnieniem: *nauka* a *technologia* (*science* vs *technology*). Zgodnie z tym podziałem *biologia obliczeniowa opracowuje modele ilościowe* zagadnień biologicznych, często bardzo teoretyczne, bazujące na sformułowaniach matematycznych, bioinformatyka natomiast korzystając z tych modeli (lub znacznie je upraszczając) buduje algorytmy konstruujące interesujące biologów rozwiązania. Nie jest to jednak jedyne możliwe rozróżnienie między tymi dwiema dziedzinami, np. J. Xiong [43] proponuje taki podział:

„*Bioinformatyka różni się od powiązanej z nią dziedziny zwanej biologią obliczeniową, gdyż ogranicza się do analizy sekwencji, struktury oraz funkcji genów i genomów oraz odpowiadających im produktów ekspresji. Dlatego często określa*

się ją mianem molekularnej biologii obliczeniowej. Biologia obliczeniowa natomiast obejmuje wszystkie obszary biologii, które wymagają obliczeń. Na przykład w modelowaniu matematycznym ekosystemów i dynamiki populacji, w zastosowaniu teorii gier do analiz behawioralnych i rekonstrukcjach filogenetycznych wykorzystujących dane kopalne stosuje się narzędzia obliczeniowe, które nie muszą mieć związku z makrocząsteczkami biologicznymi".

Wydaje nam się jednak, że pierwszy podział lepiej pasuje do ogólnej metodologii nauk, przypomnianej niedawno na jednym z wykładów inauguracyjnych przez L. Pacholskiego [29]. Trawestując cytowaną tam anegdotę pochodzącą od prof. Jana Kaczmarka, *biologia obliczeniowa zajmuje się poznaniem tego co jest, natomiast bioinformatyka – tworzeniem tego, czego jeszcze nie ma*. Zatem przyjmiemy dalej, iż bioinformatyka rozwija narzędzia do analiz sekwencji (kwasów nukleinowych i białek), strukturalnych, funkcjonalnych oraz systemowych (por. zmodyfikowany podział dziedziny na Rys. 2, pierwotnie zaproponowany w [43]), a także narzędzia do ekstrakcji informacji przechowywanej w bazach danych.



Rys. 2. Podział bioinformatyki.

Celem bioinformatyki jest pomoc w zakresie zrozumienia funkcjonowania mechanizmów organizmów żywych poprzez budowę i wykorzystanie narzędzi ilościowych. Zastosowania tych badań znacząco wykraczają jednak poza nakreślony powyżej obszar, obejmując także wiele pokrewnych dziedzin jak biotechnologię czy medycynę, przyczyniając się np. do szybszego projektowania leków, wykonywania analiz DNA w medycynie sądowej, czy też przybliżając moment, w którym każdy z nas będzie mógł korzystać z dobrodziejstw *medycyny spersonalizowanej*.

W niniejszym studium znajdą się, choć oczywiście omówione bardzo skrótowo, powyższe aspekty dotyczące zakresu badawczego i zastosowań bioinformatyki, ze szczególnym uwzględnieniem wyników środowiska poznańskiego i perspektyw jej rozwoju na Politechnice Poznańskiej.

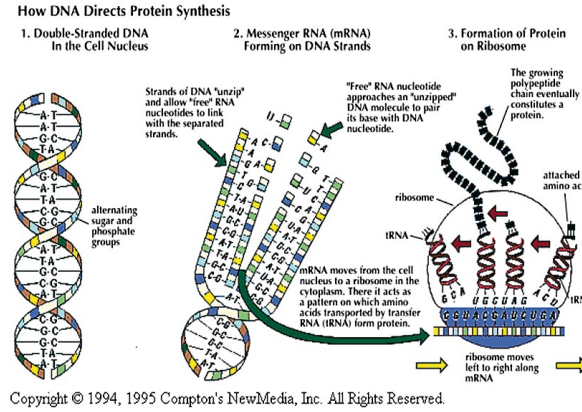
2. Biologiczne podstawy

Biorąc pod uwagę fakt, że wykład ma miejsce na Politechnice Poznańskiej, niezbędne wydaje się przypomnienie podstawowych praw i odkryć biologii. W świetle tych ostatnich można stwierdzić, że wszystkie żyjące organizmy składają się z jednej lub wielu *komórek*. Stąd jednym z najważniejszych działów biologii jest *biologia komórkowa* (a idąc dalej tym śladem *obliczeniowa biologia komórkowa*, którą niektórzy utożsamiają z bioinformatyką). Jednym z najważniejszych obiektów, którymi zajmuje się biologia są *łańcuchy DNA* (molekuły kwasu nukleinowego DNA), kodujące informację genetyczną opisującą budowę i funkcjonowanie nie tylko pojedynczych komórek, ale i całych organizmów. Należy podkreślić, że budowa komórek różnych organizmów, a jeszcze bardziej kodujących je łańcuchów DNA wykazuje zadziwiające podobieństwa. DNA jest *ciągiem nukleotydów* (dla wygody informatyk może o nich myśleć jak o literach), przybierających zaledwie cztery wartości: *A, C, G, T*. Nie analizując bliżej ich budowy, możemy powiedzieć, iż różnią się one zasadą azotową wchodzącą w ich skład. Krótkie, pojedyncze łańcuchy DNA nazywane są *oligonukleotydami*.

Cała informacja genetyczna organizmu skupiona w jednym lub kilku łańcuchach DNA, nazywana jest *genomem*. Dzięki wspomnianemu, fundamentalnemu odkryciu Watsona i Cricka [41] wiadomo, że łańcuch ten ma postać *podwójnej helisy*, a co więcej zawsze naprzeciwko A w jednym łańcuchu występuje T w drugim, a naprzeciwko C mamy G (por. Rys. 1). Znając zatem jeden z łańcuchów, drugi – komplementarny, można zawsze odtworzyć. Korzysta z tej właściwości sam organizm przy rozmnażaniu, korzystają też biolodzy w laboratoriach. Długość genomu człowieka przekracza trzy miliardy par zasad – *bp* (nukleotydów – w takich jednostkach mierzony jest genom). Warto podkreślić, że kopia genomu znajduje się w każdej komórce organizmu i w przypadku człowieka jest to ponad 400 bilionów kopii. Jedną z ciekawych właściwości pojedynczego łańcucha DNA jest jego dążenie do połączenia się w helisę z łańcuchem komplementarnym. Właściwość ta wykorzystywana bywa bardzo często w laboratoriach.

Wspomnianym badaczom zawdzięczamy też *Centralny Dogmat Biologii Molekularnej*, który określa związki miedzy DNA, a białkami będącymi budulcem organizmu, jak i elementami odpowiedzialnymi za jego działanie. Otóż, *w procesie transkrypcji* odpowiedni fragment genomu jest *przepisywany*

na łańcuch *RNA* (nietrwały, pośredniczący, pojedynczy łańcuch innego kwasu nukleinowego), który przechodząc następnie do rybosomu jest tłumaczony (*podlega translacji*) na odpowiadający łańcuch białkowy (por. Rys. 3, gdzie przedstawiono przebieg procesów transkrypcji i translacji). Należy podkreślić, że fragment łańcucha DNA odpowiadający jednemu łańcuchowi białka nosi nazwę genu. Warto dodać, że trzem nukleotydom odpowiada jeden aminokwas (element budowy białka), chociaż przypisanie to (nazywane *kodem genetycznym*) nie jest wzajemnie jednoznaczne.



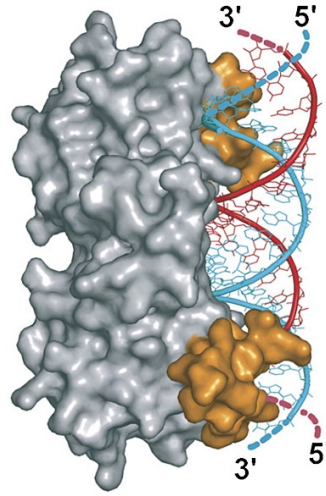
Rys. 3. Centralny Dogmat Biologii Molekularnej.

Zadziwiające jest, że powyższy schemat odpowiada w zasadzie wszystkim organizmom żywym, a stosowany kod genetyczny w dużej mierze jest taki sam w większości przypadków. Powyższy schemat, powszechnie przyjmowany jeszcze kilkanaście lat temu, zakładał, że tylko ok. 5% genomu DNA zawierało *informację kodującą* białko. Resztą było tzw. *DNA śmieciowe*.

Komputerową metaforę genetyczną tego modelu można przedstawić, przyjmując, że DNA to *pamięć długookresowa* (np. dyskowa), RNA to pamięć krótkookresowa (RAM, lub pamięć notatnikowa), a białka to *procesory*, wykonujące w systemie różne programy.

Rozwój biologii pokazał jednak, że powyższy schemat jest zdecydowanie za prosty (trudno było przypuścić, że Natura marnotrawi 95% genomów organizmów życiowych). Nowe teorie zakładają znacząco większy udział małych łańcuchów RNA (transkrybowanych z „nieopisanych" dotychczas fragmentów genomu) we wszystkich procesach organizmów żywych. W naszej *komputerowej metaforze*, RNA spełniałoby zatem (oprócz pamięci notatnikowej) także rolę *specjalizowanych mikroprocesorów* wykonujących różne operacje pomocnicze w systemie.

Powyższa metafora winna przybliżyć informatykom (a ogólniej ludziom techniki) zasady funkcjonowania organizmów żywych. Trzeba jednakże podkreślić, że jest ona bardzo uproszczona, bo działanie systemów komputerowych jest z zasady linearne, tzn. zakodowane ciągi znaków są odczytywane i wykonywany jest odpowiedni program. I chociaż w działaniu organizmów żywych można się dopatrzyć elementów takiego typu działań (same procesy transkrypcji i translacji), to jednak wiele funkcji spełnianych zwłaszcza przez łańcuchy białkowe i RNA zależy od ich kształtu przestrzennego. Analiza zatem i predykcja funkcji poszczególnych elementów takiego systemu jest znacząco trudniejsza od analizy systemów informatycznych. (Por. Rys. 4, na którym przedstawiono przykład dokowania przestrzennego domeny RNazy IIIb białka Dicer i cząsteczki dwuniciowego RNA (dsRNA). Rybonukleaza o aktywności RNazy III (zwana Dicer w komórkach ludzkich) umożliwia degradację dwuniciowych, liniowych fragmentów RNA do siRNA.)



Kompleks domena RNazy IIIb/dsRNA

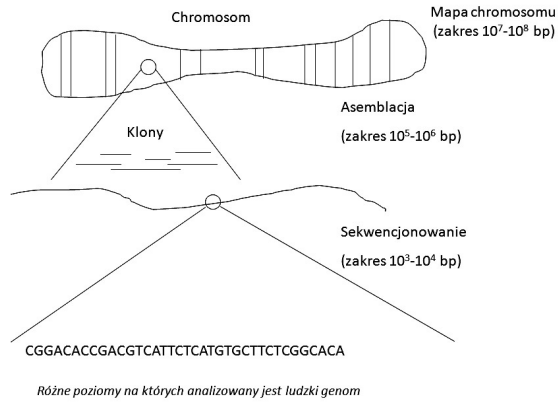
D. Takeshita, S. Zenno, W.C. Lee, K. Nagata, K. Saigo, M. Tanokura Homodimeric Structure and Double-stranded RNA Cleavage Activity of the C-terminal RNase III Domain of Human Dicer, *J. Mol. Biol.*, 374, 106-120 (2007)

Rys. 4. Przykład dokowania przestrzennego domeny RNazy IIIb białka Dicer i cząsteczki dwuniciowego RNA (dsRNA).

W dalszej części naszego opracowania dla ilustracji zastosowań bioinformatyki w biologii, omówimy proces sekwencjonowania (*czytania*) *genomów DNA*, a także zagadnienia przewidywania struktur przestrzennych molekuł. Pokazany zostanie także wpływ bioinformatyki na zagadnienia projektowania leków i problemy medycyny spersonalizowanej.

3. Czytanie genomów DNA

Historycznie rzecz biorąc, czytanie genomu DNA dokonywane było na 3 poziomach [40], [31], [37], [25]: *sekwencjonowanie (sequencing)*, *asemblacja (assembling) i mapowanie (mapping)*. Wzajemne zależności przedstawiono schematycznie na Rys. 5.

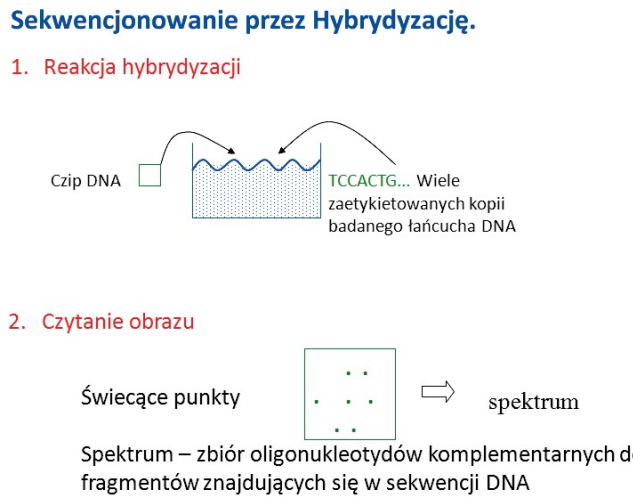


Rys. 5. Trzy poziomy odczytu genomu.

Oczywiste jest, że długiego fragmentu genomu (zwykle mówimy tak o *chromosomach* o długości rzędu 10^8bp, na które podzielone są genomy) nie można bezpośrednio odczytać. Proces ten zwykle przebiegał na trzech wspomnianych poziomach i bardzo przypominał informatyczną procedurę *rekurencyjną*. W etapie pierwszym – *mapowaniu*, dzielono chromosom na mniejsze fragmenty o długości ok. miliona par zasad (bp) za pomocą enzymów restrykcyjnych, tracąc oczywiście informację o kolejności występowania poszczególnych fragmentów. W kolejnym etapie – *asemblacji*, powyższe fragmenty ponownie były dzielone na mniejsze, tym razem przy wykorzystaniu *techniki śrutowania (shotgun technique)*. W zależności od stosowanej później metody odczytu tych najmniejszych fragmentów – *sekwencjonowania*, długość tych ostatnich wahała się od 1000 bp dla *sekwencjonowania na żelu*, poprzez ok. 500 bp dla *sekwencjonowania przez hybrydyzację (SBH)*, do 50 bp w przypadku najnowszych *sekwenatorów* NGS. Składanie genomu odbywało (odbywa) się w odwrotnej kolejności (jak już wspomnieliśmy nieomal tak jak przy rekurencji): zsekwencjonowane (przeczytane) małe fragmenty DNA łączy się w większe na etapie asemblacji, wykorzystując w tym celu *podejścia grafowe* i algorytmy heurystyczne, otrzymane zaś duże fragmenty mapuje się (składa) w całe chromosomy, rozwiązując bardzo ciekawe problemy z teorii liczb.

Omówimy poniżej metodę *sekwencjonowania przez hybrydyzację (sequencing by hybridization) – SBH*, gdyż jest ona związana z badaniami ośrodka poznańskiego (W. Markiewicz), a ponadto wymaga sporego wkładu bioinformatyki i stosowane tu podejścia można w dużej mierze wykorzystać na etapie asemblacji. (Ze względu na pewne trudności związane z techniczną stroną tego pomysłu, metoda ta nie weszła do szerokiej praktyki laboratoryjnej mimo jej opatentowania [38] i [10].)

Metoda SBH składa się z dwóch etapów. W etapie pierwszym – biochemicznym, wykorzystuje się właściwość pojedynczej nici DNA do przyłączenia nici komplementarnej w celu utworzenia struktury helisy. Pojedynczą nić badanego (nieznanego) łańcucha DNA (o długości w przypadku SBH ok. 500 bp) uzyskuje się, rozdzielając helisę poprzez stworzenie warunków umożliwiających rozpad (np. wysoka temperatura, rozpuszczalnik organiczny o dużej sile jonowej). Następnie do roztworu wprowadza się *badany pojedynczy łańcuch DNA* o długości *n* oraz pełną *bibliotekę oligonukleotydów* o długości *l* (wszystkie oligonukleotydy – *oligomery*, *l-mery* o tej długości). Bibliotekę oligomerów można zrealizować w postaci *czipa*, czyli *matrycy l-merów*, do których przyłącza się badaną nić. Te l-mery, które przyłączą się do badanego łańcucha tworzą *spektrum*, które jest podstawą jego rekonstrukcji. Dzięki reakcji polimerazy (PCR) można barwnik fluorescencyjny dodać tylko do jednej nici (takiej samej w każdym przypadku) badanego łańcucha DNA [1], [17]. Schematyczny przebieg eksperymentu pokazano na Rys. 6.



Rys. 6. Przebieg eksperymentu hybrydyzacyjnego w metodzie SBH.

Załóżmy teraz, że udało się przeprowadzić eksperyment biochemiczny bezbłędnie, a powstałe *spektrum* jest *idealne*. (Z punktu widzenia biologii jest to mało prawdopodobne, ale taka idealizacja jest niezbędna dla właściwego zamodelowania informatycznego całego procesu.) W celu ilustracji fazy biochemicznej SBH rozpatrzmy Przykład 1.

Przykład 1. Załóżmy, że oryginalna sekwencja, którą chcemy zsekwencjonować ma postać ACTCTGG, oraz, że udało się przeprowadzić eksperyment bez błędu, odczytując tę sekwencję z macierzy czipowej zawierającej 3-mery. Uzyskamy zatem następujące spektrum S={ACT,CTC,CTG,TCT,TGG}, co wynika z następującego wyniku hybrydyzacji (tworzenia helis częściowych z trójkami macierzy):

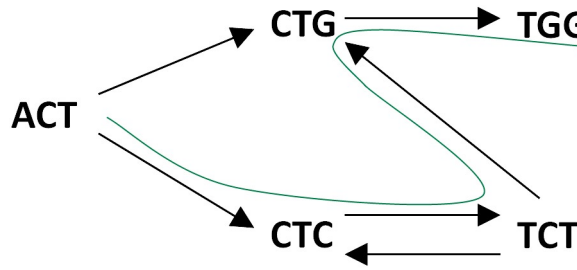
```
ACTCTGG

TGA                ACT
 GAG                CTC
  AGA    }  ⇒       TCT
   GAC                CTG
    ACC                TGG
                        □
```

Spektrum S nie daje w ogólności informacji o kolejności występowania oligonukleotydów w odczytywanym łańcuchu. I tu pomaga bioinformatyka, która akurat w tym przypadku wykorzystuje teorię grafów. Pierwszy algorytm rozwiązujący powyższy przypadek bioinformatycznej fazy SBH podał Lysov ze współpracownikami [26]. Algorytm ten jest następujący.

Algorytm Lysowa

1. Skonstruuj graf *H*, w którym wierzchołki odpowiadają *l-merom* ze *spektrum*. Dwa wierzcholki są połączone łukiem skierowanym wtedy i tylko wtedy gdy *l*-1 nukleotydów końcowych pierwszego *l-meru* jest dokładnie równych *l*-1 początkowym nukleotydom drugiego *l-meru*.
2. Znajdź w grafie *H drogę Hamiltona*, to jest drogę przechodzącą dokładnie jeden raz przez każdy z wierzchołków *H*.

Przykład zastosowania tego algorytmu pokazano na Rys. 7.



Rys. 7. Droga Hamiltona w grafie Lysova dla spektrum z Przykładu 1.

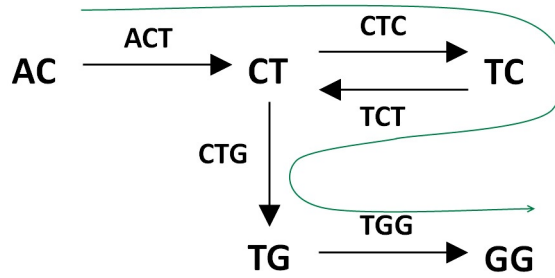
Niestety zaproponowane rozwiązanie okazało się bardzo pracochłonne, bo problem drogi Hamiltona w grafie skierowanym jest NP-trudny [20], [3], zatem nie istnieje algorytm wielomianowy rozwiązujący ten problem grafowy.

Jednak już w następnym roku, Pevzner zaproponował znacząco szybsze podejście [30].

Algorytm Pevznera

1. Skonstruuj graf G, w którym łuki skierowane odpowiadają *l-merom*, łącząc każdorazowo wierzchołek odpowiadający *l*-1 nukleotydom początkowym z wierzchołkiem odpowiadającym *l*-1 nukleotydom końcowym.
2. Znajdź w grafie G *drogę Eulera*, to jest drogę przechodzącą dokładnie jeden raz przez każdy z łuków G.

Zastosowanie powyższego algorytmu do spektrum z Przykładu 1 pokazano na Rys. 8.



Rys. 8. Graf Pevznera dla spektrum z Przykładu 1.

Co ciekawe okazało się, że powyższy problem można rozwiązać w czasie wielomianowym, gdyż taka jest złożoność znalezienia drogi Eulera w grafie skierowanym. Rodziło to dalsze pytania, bo z jednej strony problem grafowy odpowiadający SBH był *NP-trudny*, z drugiej zaś łatwy obliczeniowo, co mogło prowadzić do zaskakujących wniosków [20], [3]. (W swej oryginalnej pracy

Pevzner nie wytłumaczył tego faktu.) *Problem SBH* w wersji idealnej był zatem bardzo interesujący z punktu widzenia teorii złożoności i grafów, a jego nieidealne uogólnienie – także dla biologii (bo przypomnijmy, że zarówno podejście Lysova jak i Pevznera dotyczyły idealnego przypadku, trudnego do uzasadnienia biologicznie). Oba te kierunki badawcze związane były następnie w dużej mierze z ośrodkiem poznańskim.

Problem teoretyczny udało się rozwiązać w 1999 r. [13]. W pracy tej pokazano równoważność obu modeli, wykorzystując w tym celu pojęcie *skierowanych grafów liniowych*. Mówiąc bardziej formalnie, w skierowanym grafie liniowym H (model Lysowa) istnieje droga Hamiltona wtedy i tylko wtedy, gdy w jego grafie oryginalnym (model Pevznera) istnieje droga Eulera. Analiza powyższych zależności zaowocowała też wprowadzeniem do literatury przedmiotu koncepcji *grafów DNA* wykorzystywanych do reprezentacji wyników idealnych eksperymentów hybrydyzacyjnych. Bardziej formalnie, są to indukowane podgrafy grafów de Bruijna [13]. Ich analizie poświęconych jest ostatnio wiele prac (por. m.in. [30]).

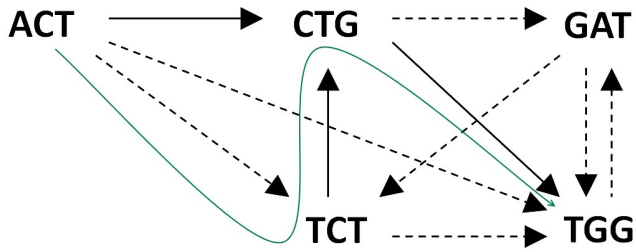
Równie interesująca okazała się analiza zagadnienia sekwencjonowania SBH z błędami wynikającymi z eksperymentu biochemicznego. Istnieją dwa typy błędów eksperymentalnych:

- *Negatywne* – w spektrum brakuje pewnych (k) oligonukleotydów w porównaniu z przypadkiem idealnym,
- *Pozytywne* – w spektrum występują oligonukleotydy nadmiarowe.

Błędy negatywne mogły mieć przy tym dwojaką naturę. Po pierwsze mogły pochodzić z samego eksperymentu i niedokładnego odczytu świecących punktów. Po drugie mogły też wynikać z powtórzeń *krótkich l-merów*, które znajdowały się w odczytywanym łańcuchu, a których to podejście nie było w stanie uwzględnić. Duży przełom w podejściu do rozwiązania problemu stanowiła praca [11]. Sformułowano w niej problem sekwencjonowania z dowolnymi błędami jako problem *komiwojażera z wyborem* (lub inaczej z premiami) – *Prize Collecting Traveling Salesman Problem (PCTSP)*, w specjalnym grafie skonstruowanym na bazie spektrum. Problem *PCTSP* różni się od *klasycznego problemu komiwojażera* [3] istnieniem *współczynników zysku (premii)* w wierzchołkach, oprócz standardowych *kosztów (odległości)* przypisanych skierowanym łukom grafu [24]. Celem jest tutaj *znalezienie drogi prostej* (niekoniecznie będącej drogą Hamiltona) *o maksymalnym zysku*, przy nieprzekroczeniu zadanego kosztu (długości drogi). W przypadku SBH, został użyty graf Lysowa, ale tym razem dopuszczono także możliwość połączenia pomiędzy wszystkimi wierzchołkami grafu (a nie tylko pomiędzy tymi, które różnią się etykietą – oligonukleotydem, przesuniętą o jeden). Koszt każdego łuku określa teraz przesunięcie pomiędzy etykietami połączonych wierzchołków, a zysk z dołączenia każdego

wierzchołka wynosi 1. W takim grafie poszukuje się drogi prostej o maksymalnym zysku (a więc maksymalnej liczbie nukleotydów ze spektrum), której sumaryczny koszt nie przekracza *n-l*, gdzie *n* jest znaną długością poszukiwanej sekwencji, a *l* – długością oligonukleotydów w bibliotece umieszczonej na czipie. Przykład 2 ilustruje tę konstrukcję.

Przykład 2. Wykorzystamy raz jeszcze spektrum z Przykładu 1, lecz tym razem wprowadzimy błędy: negatywne – oligonukleotyd *CTC* nie pojawia się w spektrum i pozytywne – pojawia się dodatkowy nukleotyd (błąd) GAT. Nasze (błędne) spektrum będzie zatem miało postać *S'*={*ACT*,*CTG*, *GAT*,*TCT*,*TGG*}. Odpowiadający mu graf pokazano na Rys. 9.



Rys. 9. Transformacja błędnego spektrum z Przykładu 2 do problemu PCTSP. Łuki ciągłe mają przypisany koszt 1, a przerywane – koszt 2. Pominięto łuki o koszcie 3.

Zauważmy, że także w tym przypadku udaje się zrekonstruować poszukiwaną sekwencję ACTCTGG pomimo obu typów błędów.

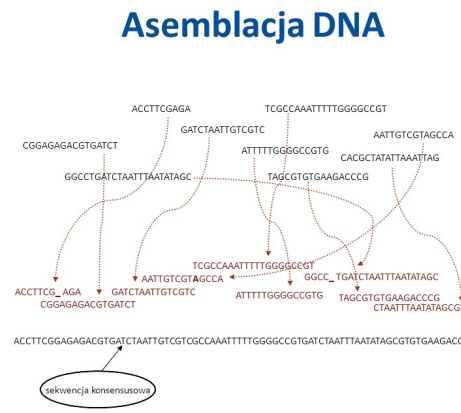
W oryginalnej pracy zaproponowano algorytm rozwiązania problemu bazujący na metodzie *podziału i ograniczeń*. Ponieważ jednak otrzymany problem *PCTSP* jest NP-trudny, więc właściwsze wydawało się rozwiązanie go za pomocą algorytmów heurystycznych. Szereg takich algorytmów zaproponowano w późniejszym czasie, między innymi w pracach [5], [12], [14], [6]. W efekcie rozwiązano rzeczywisty problem biologiczny (z błędami) uzyskując dobrą dokładność rekonstruowanych sekwencji, za wyjątkiem jednego przypadku – błędów negatywnych pochodzących z powtórzeń. Biblioteka *l-merów* nie była bowiem w stanie przekazać liczby świecących, powtarzających się oligonukleotydów. Ponieważ liczne powtórzenia są immanentną cechą sekwencji DNA, więc przesądziło to o stosunkowo niewielkim wykorzystaniu praktycznym tego podejścia.

Ostatnio, wraz z rozwojem *Projektu Genomu Człowieka* znacząco rozwinął się przemysł produkujący *sekwenatory DNA nowej generacji* (Next Generation Sequencers). Znacząco przyśpieszyły one sekwencjonowanie DNA, eliminując

równocześnie konieczność korzystania z obliczeń algorytmicznych na tym etapie. I tak na przykład sekwenator firmy Illumina umożliwia równoległe odczytanie wielu milionów sekwencji. Sekwencjonowanie przez syntezę odbywa się poprzez dołączanie kolejnych nukleotydów do drugiej komplementarnej nici DNA. Każdy z czterech różnych nukleotydów generuje inny sygnał świetlny rejestrowany przez kamerę. Analizując kolejne obrazy z kamery, można odczytać każdą sekwencję. Długość sekwencjonowanych fragmentów DNA wynosi 50-200 nukleotydów, a w ciągu kilku dni można zsekwencjonować cały genom np. człowieka. (Tu należy podkreślić, że uzyskuje się tylko ogromną liczbę tych niewielkich sekwencji składowych, podobnie jednak jak uprzednio traci się informację o kolejności ich występowania w genomie.) Na Rys. 10 przedstawiono przykładowy sekwenator Illuminy – *Genome Analyzer II*.

Rys. 10. Sekwenator firma Illumina – Genome Analyzer II.

Paradoksalnie okazało się, że chociaż podejście do sekwencjonowania przez hybrydyzację (SBH), nie znalazło dużego zastosowania w praktyce, to jego rozbudowana i znacznie ulepszona część algorytmiczna (bioinformatyczna) jest podstawowym elementem *asemblacji*, czyli składania małych sekwencji DNA (uzyskanych przez sekwenatory NGS) w sekwencje genomowe. Schematyczną ilustrację procesu asemblacji pokazano na Rys. 11.



Rys. 11. Ilustracja procesu asemblacji. 30.08.2011.

Jak łatwo sobie wyobrazić, *proces asemblacji* jest dużo bardziej skomplikowany (mimo formalnie bardzo podobnej natury) od *bioinformatycznej fazy SBH*, poprzez znacząco większą skalę procesu (terabajtowe zbiory danych – sekwencji, podlegające asemblacji) oraz dodatkowe błędy wynikające

z niedokładności procesu sekwencjonowania. (Oprócz standardowych błędów odczytu mamy tu do czynienia także np. z chimerami czy też ze zmienną orientacją łańcucha DNA lub zanieczyszczeniami pochodzącymi od innych organizmów niż aktualnie sekwencjonowany.) Mimo tych komplikacji, zespół nasz dzięki doświadczeniu zdobytemu przy okazji rozwiązywania problemu SBH, skonstruował także jeden z lepszych algorytmów asemblacji dotyczących *sekwenatora 454* [7], [8]. Prace te są aktualnie kontynuowane w ośrodku poznańskim, także dla innych typów sekwenatorów.

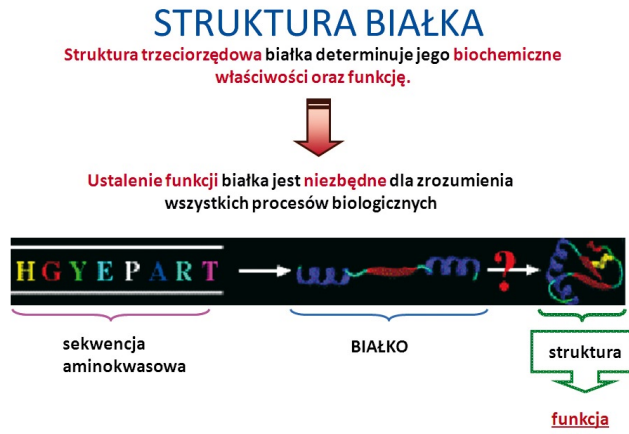
Stosunkowo niewiele mówiliśmy tu o mapowaniu. Wydaje się, że problem ten, choć bardzo interesujący z teoretycznego punktu widzenia i wykorzystujący w swej fazie algorytmicznej pewne twierdzenia z teorii liczb [40], [9], nie jest ostatnio często wykorzystywany w praktyce laboratoryjnej. Wynika to z faktu, że mapy większości genomów już istnieją, a moc nowoczesnych sekwenatorów jest bardzo duża. Warto jednakże nadmienić, że część problemów teoretycznych związanych z problemem mapowania jest nadal otwarta [4].

Należy w tym momencie wspomnieć o kolejnej fazie *czytania genomów DNA*, jak byśmy mogli powiedzieć – *czytania ze zrozumieniem*. Otóż, wiadomo już, że genomy różnych przedstawicieli danego gatunku (np. ludzi) różnią się między sobą. Te często nieznaczne różnice w sekwencji DNA (*SNP – single nucleotide polymorphism*) mogą prowadzić do znaczących różnic w odporności na choroby, czy też wręcz do wystąpienia groźnych chorób genetycznych. Aby stwierdzić występowanie określonego polimorfizmu, genom danej osoby należy po zsekwencjonowaniu interesujących fragmentów porównać z genomem referencyjnym. Czyni się tak często, wykorzystując *macierze DNA – czipy*, o których mówiliśmy w kontekście SBH.

Omówione w tym rozdziale podstawowe problemy bioinformatyczne dotyczyły łańcuchów DNA i ich analizy, czyli można by powiedzieć bioinformatyki dotyczącej struktur pierwszorzędowych, linearnych. W funkcjonowaniu organizmów żywych dużą rolę odgrywają też struktury przestrzenne biomolekuł – zwłaszcza białek i RNA; zależą od tego wykonywane przez nie funkcje. Omówimy bardzo krótko to zagadnienie w następnym rozdziale.

4. Struktury przestrzenne białek i kwasów nukleinowych – analiza, predykcja

Jak wspomniano, struktury przestrzenne biomolekuł odgrywają decydującą rolę w sprawowanych przez nie funkcjach. Zatem analiza i określenie (znajdowanie) struktur drugo- i trzeciorzędowych odgrywa ogromną rolę w biologii. Można ująć to zagadnienie tak jak na Rys. 12 zrobiono to w odniesieniu do białek.

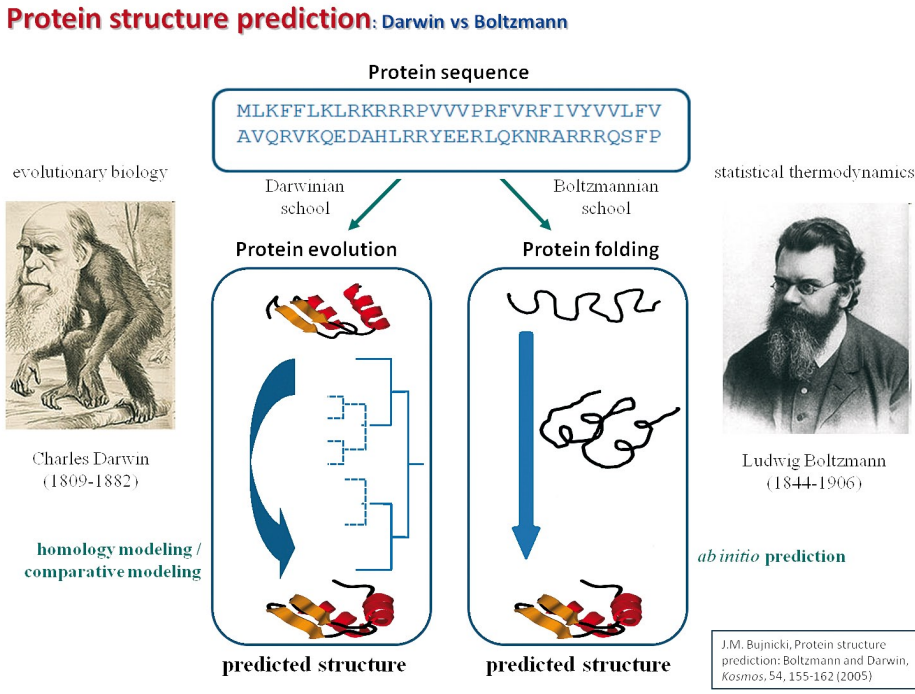


Rys. 12. Struktura trzeciorzędowa białka.

A zatem droga do poznania funkcji białka (jak wspomnieliśmy pełni ono rolę procesora w organizmie) wiedzie przez odkrycie jego struktury. (Ponieważ roli RNA wcześniej nie doceniono, więc nie badano zbyt intensywnie jego struktury przestrzennej.) W biologii strukturę przestrzenną białek najczęściej ustala się, wykorzystując jedną z dwóch technik doświadczalnych – albo *krystalografię rentgenowską* albo *spektroskopię magnetycznego rezonansu jądrowego (NMR)*. Techniki te są jednak bardzo kosztowne i wymagają długiego czasu, nic zatem dziwnego, że po parudziesięciu latach ich stosowania, mimo zaangażowania wielu zespołów badawczych z całego świata, udało się odnaleźć zaledwie 70019 struktur (zdeponowane w *Protein Data Bank – PDB* [28]; stan w dniu 30.08.2011 r.). Biorąc pod uwagę, że łańcuchów białkowych (struktura pierwszorzędowa) rozpoznano już kilkaset tysięcy, widać wyraźnie, że samo podejście eksperymentalne tu nie wystarczy. Z pomocą przychodzi bioinformatyka strukturalna (por. Rys. 2), która oferuje alternatywne sposoby określenia struktur białek.

Powyższe sposoby można przedstawić schematycznie tak jak to uczyniono na Rys. 13 (por. także [15] i [39]), prawdziwym także w przypadku struktur RNA.

Otóż, podejścia te można sklasyfikować jako metody *ab initio* i metody *modelowania homologicznego (porównawczego)*. Te pierwsze (wywodzące się w jakiejś mierze od prac Boltzmanna) polegają na stworzeniu pełnoatomowego modelu białka na podstawie sekwencji, bez korzystania z wiedzy o innych strukturach białkowych. W tym celu wykorzystuje się tutaj prawa fizykochemiczne określające zwijanie się łańcuchów białek. Nie są one jednak w wystarczającym stopniu zbadane, stąd też uzyskiwane rezultaty nie są najwyższej jakości. Stosunkowo najlepsze wyniki uzyskuje tutaj grupa D. Bakera [2],

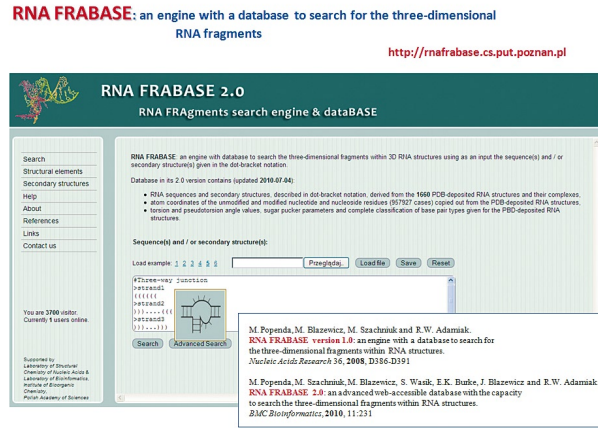


Rys. 13. Predykcja przestrzennych struktur białkowych.

a program Robetta bazujący na tych pracach jest dostępny w sieci [35]. Druga grupa metod przewiduje struktury analizowanych, nowych łańcuchów białek na podstawie podobieństwa do białek, których struktury są znane (a więc niejako wykorzystuje prawa ewolucji Darwina). Wychodzi się tu z założenia, że jeżeli dwa białka wykazują duże podobieństwo w zakresie swoich sekwencji, to podobne winny być również ich struktury trzeciorzędowe. Budowa modelu przebiega w 5 zasadniczych etapach: znajdowanie szablonu w bazie danych, przyrównanie sekwencji białka analizowanego i szablonu, budowa kręgosłupa modelu białka, dodawanie pętli i łańcuchów bocznych, optymalizacja modelu. To podejście wykorzystuje w bardzo dużym stopniu metody badań operacyjnych, w tym w szczególności: sieci neuronowe, programowanie dynamiczne i różne metody klastryzacji. Warto podkreślić, że duże sukcesy odnieśli tutaj naukowcy z Polski, w szczególności: A. Koliński, K. Ginalski, J. Bujnicki i J. Rychlewski [23], [22], [21] (por. praca przeglądowa [15]).

Sytuacja w dziedzinie analizy struktur RNA wygląda dużo gorzej niż w przypadku białek: jak dotychczas określono i zdeponowano w PDB tylko 887 struktur trzeciorzędowych. Wynika to ze stosunkowo krótkiego okresu,

w którym je analizowano jak i dużo bardziej nietrwałej struktury trzeciorzędowej RNA, którą trudno zbadać eksperymentalnie. W zasadzie wszystkie podejścia są tu bardzo podobne do omówionych w przypadku białek, lecz ze względu na większy stopień złożoności samego problemu nie ma zbyt wielu metod. Jednak i tutaj prace autorów polskich należą do wyróżniających się w dziedzinie modelowania homologicznego, a w szczególności warte podkreślenia są prace zespołów J. Bujnickiego i R. Adamiaka [36], [33], [34]. Przykładowo na Rys. 14 przedstawiono stronę tytułową serwisu WWW, w którym możliwe jest odszukanie fragmentów trzeciorzędowej struktury analizowanego łańcucha RNA, co jak wspomniano powyżej jest jednym z etapów modelowania homologicznego trzeciorzędowej struktury całego łańcucha RNA.



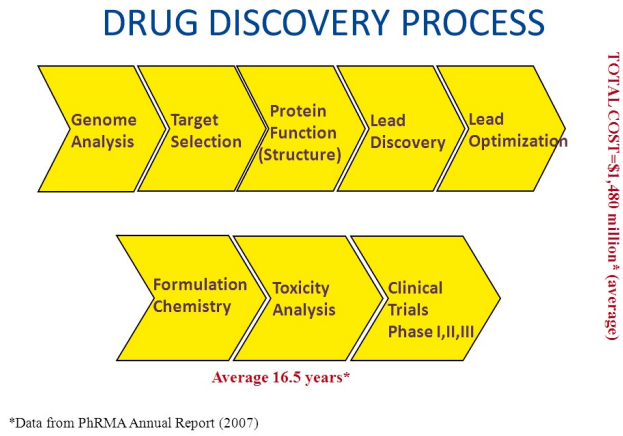
Rys. 14. RNAFrabase – serwis WWW.

Omówione pokrótce w tym rozdziale problemy to jedne z najważniejszych zagadnień, którymi aktualnie zajmuje się bioinformatyka i biologia obliczeniowa. Od ich poprawnego rozwiązania zależy w dużej mierze zrozumienie funkcji spełnianych przez określone białka i łańcuchy RNA. W efekcie proces ten przyczyni się do postępu w medycynie, a w szczególności do przyśpieszenia zagadnień projektowania leków i diagnostyki medycznej prowadzącej do medycyny spersonalizowanej.

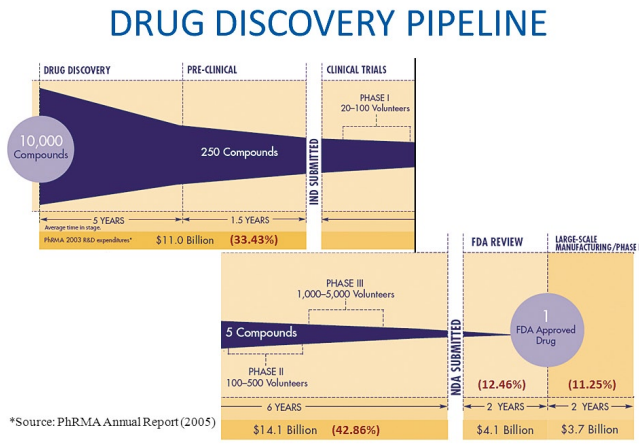
5. Wpływ bioinformatyki na medycynę

W rozdziale tym omówimy wpływ bioinformatyki na rozwój medycyny, a w szczególności na proces projektowania nowych leków i na diagnostykę medyczną. Rozpatrzmy w tym celu oficjalne dane dotyczące procesu projektowania nowych leków, opublikowane przez Stowarzyszenie PhRMA – The

Pharmaceutical Research and Manufacturers of America, skupiające najważniejsze firmy farmaceutyczne w Stanach Zjednoczonych (por. Rys. 15 i 16).



Rys. 15. Proces projektowania nowego leku.



Rys. 16. Fazy odkrycia i testowania leków w firmach zrzeszonych w PhRMA.

Jak wynika z przedstawionych rysunków projektowanie i zatwierdzanie jednego leku przy jasno określonym celu badawczym zajmowało średnio ok. 16,5 roku i kosztowało ok. 1,5 mld $. Nakłady na *rozwój i badania (R&D)* wszystkich firm skupionych w PhRMA wynosiły w 2003 r. ok. 33 mld $, natomiast w roku 2010 wyniosły ok. 49 mld $. Są to zatem kwoty bardzo

znaczące. Analizując przebieg procesu projektowania leków (Rys. 15) widzimy, że bardzo dużą część czasu (i finansów) pochłaniają analizy związane z genomiką i bioinformatyką strukturalną. W szczególności kluczowe znaczenie ma tu rozpoznanie trzeciorzędowej struktury i funkcji określonych białek. Każde przyśpieszenie tego procesu jest bezcenne, gdyż może owocować opracowaniem leku na choroby trapiące ludzkość od zarania dziejów. Stąd wynika wielka waga, jaką przykładają firmy farmaceutyczne do rozwoju badań i dydaktyki w wybranych uczelniach w dziedzinie bioinformatyki i genomiki.

Rozwój tych badań ma oczywiście znaczenie dużo większe niż tylko przyśpieszenie procesu opracowywania nowych leków. Badania te przyczyniają się bowiem również do lepszego zrozumienia istoty funkcjonowania organizmu człowieka, sprawiając, iż możliwa staje się prewencja medyczna prowadząca w kolejnym etapie do tzw. *medycyny spersonalizowanej* [16].

Jak już wspomnieliśmy, genetycznie wszyscy ludzie się różnią. Dzięki poznaniu naszego genomu, a w zasadzie wspomnianych *„punktowych różnic" (SNP)* w DNA, będzie można lepiej sterować przebiegiem procesu leczenia, a nawet efektywniej zapobiegać chorobom. W swojej znakomitej książce Francis Collins (były szef *Projektu Genomu Człowieka*; aktualnie dyrektor *NIH – National Institute of Health*) daje szereg przykładów skuteczności terapii uwzględniającej znajomość genomu człowieka [16]. Powstaje nowy paradygmat w medycynie, który ujmują poniższe punkty:

- *Każdy człowiek jest genetycznie inny i ma inne predyspozycje do zapadania na różne choroby. Należy zatem dostosować leczenie do konkretnego pacjenta (jego genomu). Stosunkowo niegroźne pojedyncze wady genetyczne mogą zostać wzmocnione i prowadzić do ciężkich chorób, gdy nałożą się na nie inne genetyczne mutacje i/lub zagrożenia środowiskowe.*
- *Możliwa będzie „wspólna terapia" dla wielu chorób naraz; np. niektóre leki na raka mogą leczyć także artretyzm lub chorobę Alzheimera.*

W ogólności dobrze znana jest zasada, że lepiej zapobiegać chorobom, niż je później leczyć. Niestety, nawet w Stanach Zjednoczonych tylko niewielka część z 2 bilionów $ przeznaczonych rocznie na lecznictwo, trafia na prewencje medyczną, a leczenie rozpoczyna się dopiero po wystąpieniu symptomów (często zaawansowanych) choroby. Aby przeciwdziałać tej tendencji, uruchomiono w USA specjalny program prewencji medycznej – *Family Health History Initiative* [18], który wraz z możliwościami poznania własnego genomu DNA (usługa oferowana także przez firmy prywatne) daje pacjentom i lekarzom znaczące korzyści i możliwość wyboru najlepszej strategii leczenia choroby, a często także unikania samego zachorowania.

Powyższe tendencje znajdują także wyraz w badaniach i dydaktyce prowadzonych na Politechnice Poznańskiej.

6. Europejskie Centrum Bioinformatyki i Genomiki

Europejskie Centrum Bioinformatyki i Genomiki (ECBiG) to unikalna na terenie Wielkopolski jednostka badawczo-rozwojowa, powstała na bazie konsorcjum zawiązanego pomiędzy Politechniką Poznańską a Instytutem Chemii Bioorganicznej PAN w Poznaniu, jako następstwo współpracy pomiędzy tymi jednostkami podjętej w 2004 roku w ramach *Centrum Doskonałości CENAT ICHB* (ang. *Center of Excellence for Nucleic Acid-based Technologies*). Wówczas to stworzone zostało pierwsze w Polsce centrum dysponujące odpowiednim przygotowaniem umożliwiającym wprowadzenie na terenie naszego kraju nowatorskich technik genomicznych (np. projektowanie, wykonywanie i analiza mikromacierzy DNA).

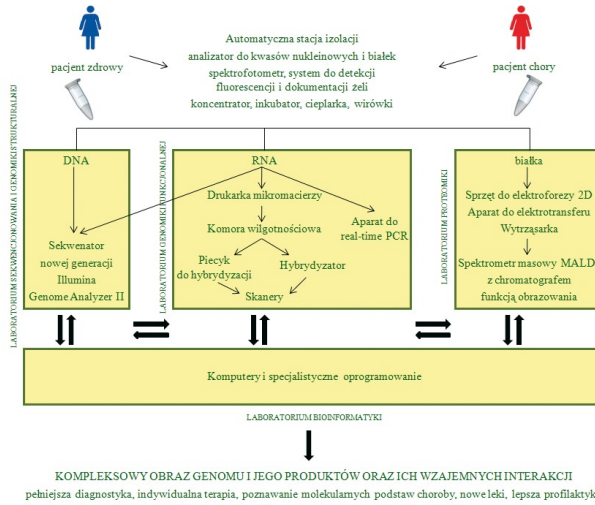
Idea stworzenia centrum łączącego badania genomiczne i bioinformatyczne, jest konsekwentnie realizowana już od kilku lat. Dzięki grantom wspólnie zdobytym przez IChB PAN (zespoły prof. prof. W. Markiewicza, M. Figlerowicza, R. Adamiaka) oraz Politechnikę Poznańską została zakupiona część sprzętu niezbędnego dla zainicjowania działalności *Europejskiego Centrum Bioinformatyki i Genomiki*. Szczególnie cennym osiągnięciem organizacyjnym było skompletowanie prawie pełnego zestawu urządzeń i programów komputerowych potrzebnych do samodzielnego projektowania, wytwarzania i analizy mikromacierzy DNA, systemu do rozdziału i analizy białek metodą elektroforezy dwukierunkowej oraz zestawu aparatów umożliwiających analizę wpływu ekspresji wybranych genów lub trans genów na metabolizm komórkowy (chromatograf cieczowy, chromatograf gazowy, spektrometr mas typu MALDI TOF wraz z reflektronem, zwiększającym rozdzielczość pomiaru). Coraz szerszy zakres badań centrum oraz olbrzymie zainteresowanie współpracą ze strony licznych ośrodków badawczych w Polsce i za granicą wiążą się jednak z koniecznością jeszcze większego zacieśnienia współpracy biologów z bioinformatykami, jak również dokonania dalszych istotnych inwestycji aparaturowych (w ramach grantu aparaturowego POIG 2.2.)

Głównym zadaniem ECBiG jest prowadzenie badań podstawowych w zakresie bioinformatyki, genomiki strukturalnej, genomiki funkcjonalnej i proteomiki, rozwój i wdrażanie nowych oraz już istniejących metod badawczych, jak i praktyczne wykorzystanie uzyskanych wyników w medycynie, biotechnologii oraz rolnictwie.

Obecnie w ramach ECBiG prowadzone są prace z następujących dziedzin: projektowanie, produkcja i praktyczne wykorzystanie mikromacierzy DNA i RNA; zastosowanie mikromacierzy DNA oraz technologii RNAi w badaniach genomu ludzkiego oraz wybranych genomów roślinnych; zastosowanie metod bioinformatycznych w analizie genomów, opracowanie i implementacja algorytmów umożliwiających projektowanie i analizę mikromacierzy DNA i RNA;

obliczeniowa analiza strukturalna i sekwencyjna białek i RNA; algorytmiczna analiza sekwencjonowania i asemblacji DNA; identyfikacja nowych regulatorowych RNA, a szczególnie mikroRNA; masowa identyfikacja i analiza białek; terapeutyczne wykorzystanie technologii RNAi (siRNA, mikroRNA) [27].

Dzięki pozytywnej decyzji Rektora Politechniki Poznańskiej cała dotychczasowa infrastruktura ECBiG, jak i ta zakupiona w ramach projektu, została przeniesiona do nowo wybudowanego Centrum Wykładowego Politechniki Poznańskiej. Do końca tego roku zostaną uruchomione laboratoria sekwencjonowania i genomiki strukturalnej, proteomiki, genomiki funkcjonalnej oraz bioinformatyki. Będą one tworzyły efektywną strukturę badawczą, skupiając całą ścieżkę badań genomicznych i bioinformatycznych w jednym miejscu oraz umożliwiając powiązanie wszystkich poziomów analiz – strukturalnych, funkcjonalnych i proteomicznych. Strukturę tę w poglądowym skrócie prezentuje Rys. 17.



Rys. 17. Schemat badawczo-sprzętowy Europejskiego Centrum Bioinformatyki i Genomiki.

W perspektywie długookresowej rezultaty prowadzonych badań w ECBiG mogą pozwolić na stosowanie medycyny spersonalizowanej w leczeniu chorób nowotworowych i wirusowych. Rozbudowana baza genomiczna pozwoli w przyszłości na skuteczne dopasowanie leku do profilu genetycznego danej osoby: technologie rozwijane w ECBiG pozwolą stwierdzić, w jakim zakresie u danego człowieka genom jest uszkodzony bądź dysfunkcyjny. Pozwolą określić także, jaka konfiguracja białek odpowiada za rozwój danej jednostki chorobowej [27].

Oprócz współpracy badawczej z Instytutem Chemii Bioorganicznej w ramach ECBiG, Politechnika Poznańska rozwija także *bioinformatykę* na polu *dydaktycznym*. Rezultatem tego jest uruchomienie w październiku 2010 roku *makrokierunku bioinformatyka*, wspólnie z Uniwersytetem im. Adama Mickiewicza w Poznaniu.

Podziękowania

Autor pragnie bardzo podziękować pp. Agnieszce Rybarczyk, Marcie Szachniuk, Aleksandrze Świercz oraz Barbarze Błażewicz za pomoc w przygotowaniu wykładu i weryfikację treści.

Literatura

[1] W. Bains, G.C. Smith, A novel method for nucleic acid sequence determination, *Journal of Theoretical Biology*, **135**, 1988, pp. 303–307.

[2] D. Baker, A. Sali, Protein structure prediction and structural genomics, *Science* **294**, 2001, pp. 93–96.

[3] J. Błażewicz, *Złożoność obliczeniowa problemów kombinatorycznych*, WNT, Warszawa, 1988.

[4] J. Błażewicz, E.K. Burke, M. Kasprzak, A. Kovalev, M. Kovalyov, The Simplified Partial Digest Problem: approximation and a graph-theoretic model, *European Journal of Operational Research* **208**, 2011, pp. 142–152.

[5] J. Błażewicz, C. Oğuz, A. Świercz, J. Węglarz, DNA sequencing by hybridization via genetic search, *Operations Research* **54**, 2006, pp. 1185–1192.

[6] J. Błażewicz, P. Formanowicz, F. Guinand, M. Kasprzak, A heuristic managing errors for DNA sequencing, *Bioinformatics* **18**, 2002, pp. 652–660.

[7] J. Błażewicz, M. Kasprzak, Graph reduction and its application to DNA sequence assembly, *Bulletin of the Polish Academy of Sciences, Tech. Sci.* **56**, 2008, pp. 65–70.

[8] J. Błażewicz, M. Bryja, M. Figlerowicz, P. Gawron, M. Kasprzak, E. Kirton, D. Platt, J. Przybytek, A. Świercz, L. Szajkowski, Whole genome assembly from 454 sequencing output via modified DNA graph concept, *Computational Biology and Chemistry* **33**, 2009, pp. 224–230.

[9] J. Błażewicz , P. Formanowicz, M. Kasprzak, Selected combinatorial problems of computational biology, *European Journal of Operational Research* **161**, 2005, pp. 585–597.

[10] J. Błażewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, Method of sequencing of nucleic acids, *The Patent Office of the Republic of Poland*, Patent No. P335786, 1999.

[11] J. Błażewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, J. Węglarz, DNA sequencing with positive and negative errors, *Journal of Computational Biology* **6**, 1999, pp. 113–123.
[12] J. Błażewicz , F. Glover, M. Kasprzak, DNA sequencing – tabu search and scatter search combined, *INFORMS Journal on Computing* **16**, 2004, pp. 232–240.
[13] J. Błażewicz, A. Hertz, D. Kobler, D.de Werra, On some properties of DNA graphs, *Discrete Applied Mathematics* **98**, 1999, pp. 1–19.
[14] J. Błażewicz, M. Kasprzak, W. Kuroczycki, Hybrid genetic algorithm for DNA sequencing with errors, *Journal of Heuristcs* **8**, 2002, pp. 495–502.
[15] M. Bujnicki, Protein structure prediction: Boltzman and Darwin, *Kosmos* **54**, 2005, pp. 155–162.
[16] F.S. Collins, *The Language of Life*, Harper Perennial, New York, 2011.
[17] R. Drmanac, I. Labat, R. Crkvenjakov, An algorithm for DNA sequence generation from k-tuple word contents of the minimal number of random fragments, *Journal of Biomolecular Structure and Dynamics* **8**, 1991, pp. 1085–1102.
[18] *Family Health History*. http:// familyhistory.hhs.gov/
[19] R.E. Franklin, R.G. Gosling, Molecular configuration in sodium thymonucleate, *Nature* **171**, 1953, pp. 740–741.
[20] M.R. Garey, D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness,* W.H. Freeman, San Francisco, 1979.
[21] K. Ginalski, N.V. Grishin, A. Godzik, L. Rychlewski , Practical lessons from protein structure prediction, *Nucleic Acids Res.* **33**, 2005, pp. 1874–1899.
[22] A. Koliński, J.M. Bujnicki, Combination of fold-recognition with de novo folding and evaluation of models. 2004, http://www.forcasp.org/upload/2165.pdf
[23] A. Koliński, J. Skolnick, Reduced models of proteins and their applications, *Polymer* **45** , 2004, pp. 511–524.
[24] M. Laporte, S. Martello, The selective travelling salesman problem, *Discrete Applied Mathematics* **26**, 1990, pp. 193–207.
[25] A.M. Lesk, *Introduction to Bioinformatics*, Oxford University Press, Oxford, 2002.
[26] P.Y. Lysov, V.L. Florentiev, A.A. Khorlin, K.R. Khrapko, V.V. Shik, A.D. Mirzabekov, Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. A new method, *Doklady Akademii Nauk SSSR* **303**, 1988, pp. 1508–1511.
[27] A. Majewski, Sojusz na rzecz bioinformatyki i genomiki, Dziennik *Gazeta Prawna*, 28.02.2011.

[28] *Protein Data Bank*. http://www.pdb.org/.
[29] L. Pacholski, Czynnik ludzki w inżynierii produkcji, Wykład Inauguracyjny, Wyd. Politechniki Poznańskiej, Poznań, 2006, pp. 23–38.
[30] R. Pendavingh, P. Schuurman, G.J. Woeginger, Recognizing DNA graphs is difficult, *Discrete Applied Mathemtics* **127**, 2003, pp. 85–94.
[31] P.A. Pevzner, l-tuple DNA sequencing: Computer analysis, *Journal of Biomolecular Structure and Dynamics* **7**, 1989, pp. 63–73.
[32] P.A. Pevzner, *Computational Molecular Biology. An Algorithmic Approach*, The MIT Press, Cambridge, MA, 2000.
[33] M. Popenda, M. Błażewicz, M. Szachniuk, R.W. Adamiak, RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures, *Nucleic Acids Res.* **36**, 2008, pp. D386–D391.
[34] M. Popenda, M. Szachniuk, M. Błażewicz, S. Wasik, E.K. Burke, J. Błażewicz, R.W. Adamiak, RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures, *BMC Bioinformatics* **11**, 2010, p.231.
[35] *Robetta*. http://robetta.bakerlab.org/
[36] M. Rother, K. Rother, T. Puton, J.M. Bujnicki, ModeRNA: a tool for comparative modeling of RAIH 3D structure, *Nucleic Acids Res.* **39**, 2011, pp. 4007–4022.
[37] J. Setubal, J. Meidanis, I*ntroduction to Computational Molecular Biology*, PWS Publishing Company, Boston, 1997.
[38] E.M. Southern, *United Kingdom Patent Application* GB8810400, 1988.
[39] M. Szachniuk, RNA tertiary structures, *Proc. ICOLE'2010*, Lessach, 2010.
[40] M.S. Waterman, *Introduction to Computational Biology. Maps, Sequences and Genomes*, Chapman & Hall, London, 1995.
[41] J.D. Watson, F.H.C. Crick, Genetic implications of the structure of deoxyribonucleic acid, *Nature* **171**, 1953, pp. 964-967.
[42] M.H. Wilkins, A.R. Stokes, H.R. Wilson, Molecular structure of deoxypentose nucleic acids, *Nature* **171**, 1953, pp. 738–740.
[43] J. Xiong, *Podstawy bioinformatyki*, Wydawnictwo Uniwersytetu Warszawskiego, Warszawa, 2011.