## 1. Opis Instytutu BioInfoBank

Bioinformatyka jest bardzo ważną dziedziną, która w znaczący sposób wpływa na nowe osiągnięcia w biologii molekularnej. Mimo tego cechuje się z reguły niskimi potrzebami infrastrukturalnymi. Efektywna grupa bioinformatyczna składa się często jedynie z paru komputerów klasy PC, ewentualnie spiętych w klaster komputerowy, oraz kilku wszechstronnych i silnie zmotywowanych naukowców, łączących znajomość biologii molekularnej, medycyny oraz umiejętność tworzenia oprogramowania komputerowego.
Taką właśnie jednostką jest Instytut BioInfoBank. Wszyscy starsi pracownicy Instytutu BioInfoBank mają kilkumiesięczny a nawet kilkuletni staż za granicą i wiele kontaktów naukowych. Wszyscy z nich mają również lukratywne oferty pracy w Stanach Zjednoczonych. Mimo to zdecydowali się zrezygnować z korzyści materialnych na rzecz tworzenia niezależnej jednostki badawczej w Polsce na wzór znakomicie działających instytutów w Stanach Zjednoczonych. Kategoryzacja Instytutu BioInfoBank przyczyniłaby się do zahamowania emigracji zdolnej i wysoko wykształconej polskiej kadry naukowej, której wieloletnia edukacja była pokrywana z budżetu państwa. BioInfoBank ma na celu stworzenie warunków do pracy naukowej na poziomie porównywalnym do placówek w Stanach Zjednoczonych czy w Europie Zachodniej.
Aby osiągnąć ten cel struktura organizacyjna Instytutu musi być (i) efektywna, by umożliwić sprawną współpracę w ramach różnorodnych projektów z wieloma partnerami reprezentujących różne grupy badawcze i przedsiębiorstwa, (ii) elastyczna, by ułatwić realizacje nowych projektów, włączanie nowych partnerów i szybkie reagowanie na zmieniające się zapotrzebowanie ze strony rynku, (iii) płaska – tak by zapewnić łatwy kontakt pomiędzy laboratoriami posiadającymi specjalistyczną wiedzę i umiejętność w różnych dziedzinach. W celu realizacji tak postawionych wymagań wprowadzone została nowoczesne zasady zarządzania poprzez projekty. Dzięki temu możliwe jest zdecydowane uproszczenie przepływu informacji i decyzji, co jednocześnie sprzyja realizacji misji Instytutu: przyciąganiu, zachęcaniu do współpracy i sprzyjaniu rozwojowi utalentowanych ludzi. Tak obrana koncepcja zarządzania w naturalny sposób wpisuje się w wymagania realizacji całościowego rozwoju Instytutu poprzez szczegółowe projekty.
Pojedynczym projektem zarządza kierownik projektu, którym w naturalny sposób będzie pomysłodawca danego projektu. Kierownik samodzielnie dobierać współpracowników do realizacji projektu. Dzięki włączeniu we wspólny projekt specjalistów z różnych często dość odległych dziedzin spodziewany jest dodatkowy pozytywny efekt przezwyciężania tzw. „egoizmu branżowego". Co więcej, multidyscyplinarne kontakty kreują zazwyczaj efekt synergii otwierający zupełnie nowe możliwości wykorzystania zgromadzonej w zespole wiedzy. Ponieważ kierownikiem projektu będzie mógł zostać jego pomysłodawca a członkiem zespołu projektowego utytułowany ekspert, przełamywana zostaje formalna hierarchia. Doktoranci są kierownikami niektórych projektów realizowanych w ramach Instytutu, mimo że osoby ze znacznie większym stażem wchodzą w skład zespołu badawczego danego projektu. Sprzyja to wyzwoleniu dodatkowych pokładów innowacyjności i kreatywności i zapewnia elastyczność w realizacji różnorodnych zadań Instytutu.
Wiele opublikowanych rezultatów projektów realizowanych w ramach jednostki ukazało się w renomowanych zachodnich czasopismach, takich jak „Science", „Cell", „PNAS", "TiBS", "Journal of Biological Chemistry", „Nucleic Acid Research", "FASEB Journal", "FEBS Letters", "Proteins", "Protein Science", "Bioinformatics" i innych. Aktualnie ukazało się już kilkadziesiąt publikacji z afiliacją Instytutu BioInfoBank, mimo krótkiego okresu pracy naukowej w ramach tej jednostki. Wszystko to dzięki zapałowi i zaangażowaniu wszystkich współpracowników. Mimo krótkiej historii grupa ta jest jedną z najlepszych polskich jednostek w tej dziedzinie i zdobywa coraz większe uznanie za granicą. Jednymi z naszych głównych osiągnięć w ostatnich latach są wysokie miejsca w prestiżowych międzynarodowych „zawodach" (CASP-5, CASP-6). Udało nam się odkryć ważną funkcję jednej z domen białka produkowanego przez wirusa SARS. Publikacja na ten temat ukazała się w prawdopodobnie najbardziej prestiżowym czasopiśmie biologii molekularnej, jakim jest „Cell" („impact factor" około 30 w ostatnim roku, czyli więcej niż „Nature" czy „Science") i została zauważona przez międzynarodową agencję prasową Reuters-Health (2003-06-03) jako istotne osiągnięcie polskich naukowców. Jedno z czołowych czasopism naukowych „Science" opublikowało dwie nasze prace dotyczące analizy transkryptomu ryżu oraz racjonalnie zaprojektowanego sztucznego białka.
Poza własną działalnością badawczą, BioInfoBank udostępnia na swoich stronach internetowych kilka bardzo nowatorskich serwisów, które są wykorzystywane przez wielu naukowców z całego świata. W pierwszym numerze specjalnego wydania czasopisma „Nucleic Acid Research" poświęconego wyłącznie serwisom internetowym, serwer 3D-Jury rozwinięty i utrzymywany przez naszą jednostkę jest wymieniony na pierwszym miejscu i znalazł się na okładce.
Biorąc pod uwagę aktualne osiągnięcia jesteśmy przekonani, że jednostka ta będzie godnie reprezentować

naukę polską. Brak kategoryzacji naszej jednostce oznaczałoby brak zainteresowania ze strony Polskiego Rządu młodymi, prężnymi grupami naukowców i pogrzebałoby ich wiarę w perspektywy badań naukowych w Polsce. Z drugiej strony pozytywne rozpatrzenie naszego podania byłoby istotnym sygnałem dla młodych naukowców. Doprowadziłoby to do zwiększenia zainteresowania karierą w branży „wysokich technologii" i miałoby inspirujący wpływ na młodą kadrę naukową. Według nas, jednostki tego typu powinny stanowić istotną część krajobrazu naukowego w Polsce, dokładnie tak, jak ma to miejsce w przypadku instytucji „non for profit" w Stanach Zjednoczonych. Instytucje tego typu przyczyniają się nie tylko do tworzenia zatrudnienia dla wysoko wykształconej kadry, ale również do zwiększenia poziomu technologicznego w kraju.

## 2. Projekty naukowe realizowane przez jednostkę i finansowane przez Komisje Europejską

**MiFriend (2003-2004)**

Celem projektu jest analiza genomu i metabolizmu bakterii P. putida KT2440 oraz przystosowanie jej do celów oczyszczania środowiska. Zadania realizowane przez BioInfoBank to przede wszystkim automatyczna analiza genomu bakterii oraz szczegółowa analiza strukturalna i funkcjonalna wybranych rodzin genów / białek i udostępnienie pozyskanej wiedzy szerokiej społeczności akademickiej.

| Projekt | UE | MNiI | Suma |
|---|---|---|---|
| ELM | 500 | 350 | 850 |
| MiFriend | 650 | 455 | 1,105 |
| **5PR Suma** | 1.150 | 805 | 1,955 |
| SEPSDA | 1.185 | 711 | 1,896 |
| BioSapiens | 915 | 549 | 1,464 |
| DataGenome | 600 | 360 | 960 |
| MicrobeArray | 475 | 285 | 760 |
| GeneFun | 750 | 450 | 1,200 |
| **6PR Suma** | 3.925 | 2,355 | 6,280 |

Szacunkowe budżety projektów realizowanych od roku 2000 w Instytucie BioInfoBank finansowanych w ramach 5PR, dofinansowanych przez MNiI na poziomie 70% i w ramach 6PR, dofinansowanych przez MNiI na poziomie 60% przedstawia tabela obok. Wszystkie wartości podane w tysiącach PLN.

Dla każdego białka z proteomu P. putida KT2440 zostały zachowane struktura drugorzędowa przewidziana automatycznie oraz uszeregowania (alignmenty) sekwencji białek homologicznych znalezionych standardowymi narzędziami (PSI-Blast i RPS-Blast) w specjalnie stworzonej bazie danych. Zalignowane sekwencje zostały przepisane na profile i zapamiętane w celu późniejszego porównania ich z innymi profilami za pomocą czułych metod detekcji podobieństwa sekwencyjnego takich jak FFAS czy ORFeus. Baza danych zawiera rezultaty przewidywania struktury trzeciorzędowej białek uzyskane za pomocą wybranych metod. Baza danych umożliwia manualną adnotację struktury oraz funkcji białek bakterii. Adnotacje zawierają informacje o tym czy bazują na informacji eksperymentalnej czy na przewidywaniach. Baza danych zawiera również informacje o ekspresji genów otrzymane w ramach projektu w innych współpracujących laboratoriach. Zmiany w ekspresji są użyte do analizy funkcji białek. Baza danych umożliwia grupowanie białek na podstawie podobieństwa sekwencyjnego i strukturalnego. Baza danych jest dostępna przez Internet, aby umożliwić zdalną modyfikację informacji oraz udostępnić informacje szerokiej rzeszy użytkowników.

Geny odpowiedzialne za reakcje bakterii na stres są poddane szczegółowej analizie. Sekwencje odpowiednich białek, produktów tych genów, są użyte do poszukiwania sekwencji homologicznych, za pomocą metod PSI-Blast i RPS-Blast. Jeżeli poszukiwania nie doprowadzają do homologicznych białek o znanej strukturze albo funkcji, dalsze poszukiwania są przeprowadzone za pomocą bardziej czułych metod takich jak ORFeus w ramach systemu GRDB. W zależności od istotności statystycznej znalezionych podobieństw, przewidywania są adnotowane jako bardziej lub mniej pewne. W przypadku, kiedy analizowany gen można zaklasyfikować do ciekawej rodziny, przeprowadzona jest analiza filogenetyczna, która umożliwia lepszą ocenę funkcjonalnych możliwości białka. Przewidywanie struktury jest użyte do planowania badań eksperymentalnych.

Podobna Analiza jest przeprowadzona również dla białek i genów wchodzących w skład peryferyjnych szlaków katabolicznych bakterii. Szczególny nacisk jest położony na analizę genów odpowiedzialnych za katabolizm toksycznych związków z rodziny nitrotoluenu i styrenu.

Sekwencja każdego białka bakterii jest wysłana do serwera RPS-Blasta, który zawiera 400,000 profili rodzin białek wyciągniętych z bazy danych NR klastrowanej na poziomie 70%. Sekwencje znalezione przez serwer są zalignowane do sekwencji z bakterii i tak stworzony alignment (uszeregowanie) są użyte jako zalążek do pięciu iteracji PSI-Blasta. Sekwencje znalezione przez PSI-Blast powiększają informacje o rodzinach sekwencji białek bakterii, ktore są zapisane jako profile sekwencyjne. Profile są wykorzystane przy porównaniach dokonywanych za pomocą programów takich jak FFAS czy ORFeus. Profile są porównane do profili rodzin sklasyfikowanych w bazach COGs i Pfam. Zakładamy, że takie porównanie doprowadzi do

znalezienia najbardziej odległych zależności między rodzinami białek, które są wykrywalne aktualnymi metodami. Te zależności są użyte do inferencji funkcji. W przypadkach, kiedy taka inferencja jest niemożliwa, przeprowadzamy analiza strukturalna białek (w niektórych ciekawych przypadkach również manualnie), aby zawęzić spektrum możliwych funkcji.
Przewidywanie struktury zaczyna się od przewidywania struktury drugorzędowej. Następnie sekwencja badanego białka zostaje przewleczona (threading) przez znane struktury białek, aby dopasować ją do odpowiedniego szablonu. W przypadku, gdy dopasowanie (threading) jest niejednoznaczne, do dalszych przewidywań zostaje użyty „Meta-Server". Z powodu wysokich kosztów obliczeniowych, tego typu analiza jest wykonana dla nie więcej niż 500 rodzin. Automatycznie przewidziana struktura jest udostępniona w ramach serwisu internetowego, stworzonego wraz z innymi partnerami.

**ELM (2003-2004)**
Celem projektu jest analiza sekwencji białek eukariotycznych oraz poszukiwanie w ich sekwencjach krótkich motywów funkcjonalnych (ELM'ów). Motywy takie odgrywają kluczową rolę w procesach zachodzących pomiędzy białkami i są odpowiedzialne za sygnały przekazywane w ramach dużych sieci sygnałowych. Poznanie tych motywów umożliwi zrozumienie niektórych aspektów oddziaływania pomiędzy białkami i pomoże zrozumieć procesy zachodzące w komórkach eukariotycznych.
W ramach projektu Został rozwinięty i przetestowany filtr „ab-initio" do systemu ELM, aby rozszerzyć zbiór filtrów strukturalnych i umożliwić analizę preferencji konformacyjnych miejsc białek, dla których nie znana jest struktura ani też nie jest znana struktura dla innych homologicznych białek. Filtr jest użyty do lokalizacji miejsc proteolitycznych w białkach. Po zakończeniu pracy nad filtrem będzie on użyty do rozwoju opisu miejsc limitowanej proteolizy i będzie dodany do innych filtrów i programów. Baza danych ELM'ów (liniowych motywów eukariotycznych) jest rozszerzona o instancje motywów limitowanej proteolizy wraz z odpowiednimi opisami.
BioInfoBank jest odpowiedzialny za rozwój filtru „ab-initio" i użycie go do detekcji miejsc limitowanej proteolizy. Filtr „ab-initio" sprawdza kompatybilność sekwencji (segmentu białka) z badaną strukturą. Filtr bazuje na liście konformacji fragmentów wyciętych ze struktur białkowych. Konformacje są poklastrowane przy użyciu informacji strukturalnej i sekwencyjnej. Każdy klaster zawiera fragmenty podobne sekwencyjnie i strukturalnie. Filtr ab-initio dzieli białko na nakładające się na siebie fragmenty. Każdy fragment jest porównany z listą wcześniej przygotowanych klastrów fragmentów. Porównanie odbywa się za pomocą podobieństwa sekwencji lub podobieństwa profili sekwencji. Najbardziej podobne klastry są przypisane fragmentom. Zbiór podobnych klastrów służy do oceny preferencji konformacyjnych badanego fragmentu. Średnia konformacja oraz jej odchylenie są obliczane. Konformacje są porównane z wymogami poszczególnych badanych proteaz. Wiele miejsc proteolitycznych można opisać jako giętkie i posiadające małą liczbę zakonserwowanych aminokwasów. W przypadku kiedy jest znane tylko jedno miejsce proteolizy dla danej proteazy (tak jak w przypadku koagulacji krwi) ocena różnorodności konformacji jest niemożliwa. Jednak w innych przypadkach zakładamy, że specyfikacje strukturalne mogą być dodane do procesu filtrowania wykonywanego przez system poszukiwania ELM'ów. Filtr ab-initio pomaga innym filtrom również przy poszukiwaniu innych motywów niż motywy limitowanej proteolizy. BioInfoBank jest odpowiedzialny za wpisanie motywów limitowanej proteolizy do bazy danych ELM.

**BioSapiens (2004-2008)**
Sieć Doskonałości integruje 24 czołowe europejskie laboratoria bioinformatyczne:

**EMBL** European Bioinformatics Institute, Hinxton, Cambridge, UK
**EBI** European Molecular Biology Laboratory, Heidelberg, Germany

**GSF**
Environment and Health, Neuherberg, Munich, Germany

**ULB**
Université Libre de Bruxelles, Brussels, Belgium

**CSIC**
Consejo Superior de Investigaciones Cientificas, Madrid, Spain

**KUN**
University of Nijmegen, Nijmegen, The Nertherlands

**SIB**
Swiss Institute of Bioinformatics, Geneva, Switzerland

**DTU**
Technical University of Denmark, Lyngby, Denmark

**UH**
University of Helsinki, Helsinki, Finland

**IMAS**
Institut Municipal d'Assistència Sanitària, Barcelona, Spain

**Sanger**
Genome Research Ltd, Hinxton, Cambridge, UK

**MPII**
Max-Planck Institute for Informatics, Saarbrucken, Germany

**HUJI**
The Hebrew University of Jerusalem, Givat Ram, Israel

**Bio-Rome**
Department of Biochemical Sciences University of Rome "La Sapienza", Rome, Italy

**STO**
University of Stockholm , Stochkolm, Sweden

**UOXF.H5**
University of Oxford, Oxford, UK

**UCL**
University College London, London, UK

**UNIGE**
University of Geneva, Geneva, Switzerland

**ENZIM**
Institute of Enzymology, Hungarian Academy of Sciences, Budapest, Hungray

**UNIK**
University of Cologne, Cologne, Germany

**IP**
Institut Pasteur, Paris, France

**BIB**
BioInfoBank Institute, Poznan, Poland

**MPIMG**
Max Planck Institute for Molecular Genetics, Berlin, Germany

**Genoscope**
Genoscope, Evry, France

**UNIBO**
University of Bologna, Bologna, Italy

Celem BioSapines jest głębsze poznanie informacji zakodowanej w genomie ludzkim za pomocą nowoczesnych metod bioinformatycznych oraz danych eksperymentalnych. Sieć stworzy Europejski Wirtualny Instytut Adnotacji Gnomów. Partnerzy sieci będą zaangażowani w poprawę jakości badań bioinformatycznych w Europie poprzez skoncentrowany wysiłek, unikanie duplikowania czynności badawczych, wspólne spotkania, warsztaty i konferencje oraz skoordynowaną pracę badawczą. Ważnym aspektem programu jest wzmożenie współpracy między jednostkami obliczeniowymi i eksperymentalnymi poprzez ukierunkowany program analizy genomu skoncentrowany na istotnych problemach biologicznych. Adnotacje i Analizy stworzone przez sieć będą opublikowane i udostępnione w formie serwisów internetowych. Do tego celu zostanie wykorzystany system rozproszonej adnotacji (distributed annotation system DAS), który wykorzysta nowe osiągnięcia w zakresie „grid'ów" obliczeniowych. Celem sieci jest również stworzenie Europejskiej Szkoły Bioinformatyki. Celem szkoły jest edukacja bioinformatyków, którzy byliby w stanie wprowadzić najwyższe standardy w eksploatacji i wykorzystaniu informacji genetycznej organizmów. Szkoła będzie prowadzić kursy na wszystkich szczeblach i będzie otwarta dla wszystkich chętnych. Europejscy naukowcy są tradycyjnie bardzo aktywną grupą działającą w dziedzinie adnotacji genów i białek a Ensembl i SWISS-PROT (teraz UniProt) są najważniejszymi źródła informacji z tej branży używanymi na całym świecie. Wiele metod używanych przy adnotacji gnomów, sekwencji i struktur genów i białek oraz do analizy szlaków metabolicznych i sygnałowych zostało stworzonych w Europie. Sieć BioSapiens będzie dalej podwyższać kompetencje i konkurencyjność Europy poprzez nowe odkrycia, wzmożoną integrację, edukację oraz polepszenie i usprawnienie metod obliczeniowych. Sieć zwiększy rolę Europy w eksploatacji informacji genetycznej. Głównym celem integracji polskiej jednostki w Sieć doskonałości było umożliwienie wykorzystania ekspertyzy i unikalnych metod przewidywania struktury białek opracowanych przez Instytut BioInfoBank. Budżet całego projektu to 12 milionów Euro. Projekt realizuje następujące obszary priorytetowe: 1) zdrowie i życie, 2) technologie informacyjne.

**GeneFun (2004-2006)**

Odszyfrowanie informacji zawartej w sekwencjach genomowych, rozumiane jako poznanie funkcji genów i białek stanowi wielkie wyzwanie ery post-genomowej. Obecnie gros przyporządkowań funkcji do nowo zsekwencjonowanych genomów, uzyskuje się przy użyciu bioinformatycznych narzędzi, przewidujących funkcję genu na podstawie podobieństwa jego sekwencji do [sekwencji] innych genów o znanej funkcji. Uznanym faktem jest, iż te pierwotne, bazujące na podobieństwie sekwencji, procedury przyporządkowania funkcji są częstokroć niedokładne i podatne na błędy. Dalsze ich stosowanie bez jasnego i jawnego określenia granicy ich użyteczności prowadziłoby do nienaprawialnej propagacji błędów, która mogłaby zagrozić postępowi wiedzy biologicznej. Z drugiej strony, stają się dostępne różnorakie nowe zbiory danych

i zasobów [do ich analizy]. Dostarczają one informacji kontekstowej o biologicznej funkcji genów, głównie o fizycznych i funkcjonalnych interakcjach pomiędzy genami i białkami, jak również o całych procesach i sieciach [regulacyjnych]. Równolegle funkcjonujące światowe inicjatywy w dziedzinie genomiki strukturalnej, dostarczają o wiele lepszego obrazu motywów strukturalnych przyjmowanych przez białka, oraz interakcji pomiędzy nimi. Dostępność tych dodatkowych i nowych danych tworzy nie mającą precedensu okazję do stworzenia metod dla inkorporacji danych funkcjonalnych wyższego poziomu w ramy procesu adnotowania sekwencji [annotation pipeline]. Celem projektu GENEFUN jest odpowiedź na dwa wyżej wspomniane zagadnienia. Czynnik błędów w adnotacji sekwencji zostanie opanowany przez rozwój kryteriów oceny wiarygodności adnotacji obecnie dostępnych w bazach danych. Kryteria te zostaną wykorzystane do przydzielenia współczynników wiarygodności i włączone w przyszłe, ustandaryzowane procesy adnotacji sekwencji. W kwestii inkorporacji adnotacji wyższego poziomu w adnotacje funkcjonalne, zostaną połączone informacja o strukturze i sekwencji w celu identyfikacji nie-liniowych cech sekwencji (np. miejsc interakcji). Podobnie zintegrowane zostaną dostępne i nowo utworzone metody przewidywania funkcji na podstawie informacji kontekstowej i wyższego poziomu (architektura domenowa białek, interakcje białko-białko, kontekst genomowy jak np. uszeregowanie genów). Aby osiągnąć te cele kilka europejskich grup z dużym doświadczeniem w tworzeniu nowych metod i prowadzeniu analiz w zakresie genomiki porównawczej, bioinformatyki zorientowanej na strukturę i systemy, oraz w informatyce, podjęło współpracę z grupą eksperymentatorów z Kanady, znaną ze swych wybitnych osiągnięć na polu strukturalnej i funkcjonalnej genomiki. Oczekiwane wyniki projektu GENEFUN to: ulepszone procedury przewidywania funkcji na bazie podobieństwa sekwencji, zbiór procedur do przewidywania (w zautomatyzowany sposób) nie-liniowych cech funkcjonalnych na podstawie sekwencji i struktury 3D, oraz przetestowane pod względem efektywności procedury predykcji cech funkcjonalnych związanych z kontekstem. Duży nacisk zostanie położony na stworzenie protokołów postępowania łączących, w optymalny sposób, wyniki [analiz przy użyciu] większej liczby metod. W szczególności, zostaną stworzone i udostępnione społeczności naukowej serwery sieciowe, zarówno oparte na pojedynczych procedurach jak i ich zestawach. Narzędzia te będą popularyzowane w społeczności naukowej na drodze otwartych warsztatów i sesji treningowych. Rozwinięte narzędzia powinny stanowić ważki wkład w dzieło poprawy adnotacji in silico funkcji genów, a co za tym idzie wywrzeć istotny efekt na cały sektor biotechnologiczny, w znacznym stopniu bazujący na wspomnianych adnotacjach. Projekt realizuje następujące obszary priorytetowe: 1) zdrowie i życie, 2) ekologia, 3) technologie informacyjne.

**DataGenome (2004-2006)**

Chiralność jest kluczowym czynnikiem efektywności wielu leków, co powoduje, że coraz ważniejsza staje się produkcja poszczególnych enancjomerów związków chiralnych. Kataliza enzymatyczna w przeciwieństwie do syntezy chemicznej oferuje wysoką enancjoselektywność i regioselektywność w syntezie związków chiralnych. Dane molekularne są niebywałym źródłem enzymów do bio-katalizy. W celu pełnego wykorzystania potencjału tego źródła konieczne jest opracowanie nowych efektywnych metod ich pozyskiwania. Projekt DataGenome skupia się na wyszukiwaniu nowych enzymów z publicznych i komercyjnych genomów bakteryjnych (w szczególności nowych dehydrogenaz alkoholowych i monooksygenaz), oraz identyfikowania kluczowych aminokwasów w celu opracowania innowacyjnych procesów syntezy asymetrycznej. Projekt wykorzystuje analizę genomową, klonowanie, ekspresję, produkcję enzymów i inżynierię białkową do enzymatycznej produkcji związków chiralnych. Projekt wykorzystuje dużą ilość danych wyjściowych w celu przeszukania jak największej ilości genów uzyskując w ten sposób wysoką wydajność w wyborze enzymów będących najlepszymi kandydatami. Unikalne połączenie ekspertyzy i planowania badań umożliwi wysoką skuteczność udoskonalania nowych biokatalizatorów. Szczególny nacisk zostanie położony na efektywną analizę bioinformatyczną by zminimalizować wykorzystanie bardziej wymagającej „mokrej" analizy laboratoryjnej, a także rozwoju zoptymalizowanych wektorów ekspresyjnych w celu efektywnej ekspresji genów i produkcji enzymów. Racjonalna inżynieria białek oraz ukierunkowana ewolucja molekularna zostanie wykorzystana w celu otrzymania efektywniejszych enzymów, nowych preferencji substratowych lub rozszerzonej selektywności w kierunku określonych enancjomerów. Wybrane enzymy zostaną przetestowane w istniejących lub nowych procesach bio-katalitycznych w celu produkcji związków chiralnych o zastosowaniu terapeutycznym w leczeniu schorzeń takich jak AIDS, nowotwory, czy choroba Alzheimera. Głównym wkładem jednostki polskiej bioinformatycznej jest opracowanie i zastosowanie nowych metod do przewidywania specyficzności substratowej enzymów. Projekt realizuje następujące obszary priorytetowe: 1) nowe materiały, 2) ekologia, 3) technologie informacyjne.

**Sepsda (2005-2007)**
Posiadanie specyficznych i efektywnych przeciwwirusowych lekarstwa oraz nowoczesnych narzędzi diagnostycznych jest konieczne do walki z ludzkim koronawirusem wywołującym nietypowe zapalenie płuc (ang. Severe Acute Respiratory Syndrome, SARS). Chińsko-Europejski projekt dotyczący przeciwwirusowego leczenia oraz diagnozowania choroby SARS (ang. The Sino-European Project on SARS Diagnostics and Antivirals, SEPSDA) jest zintegrowanym przedsięwzięciem stosującym nowoczesną biotechnologię, w celu stworzenia potencjalnych przeciwwirusowych lekarstw i opracowania udoskonalonych metod diagnostycznych. Projekt SEPSDA połączył siły i umiejętności wiodących naukowców z Niemiec, Danii, Polski i Chin, którzy prowadząc badania nad SARS uzyskali znakomite, opublikowane osiągnięcia w dziedzinie biologii molekularnej, a dotyczące ludzkiego koronawirusa. Kilka z opracowanych związków o potencjalnych właściwościach leczniczych, jak również pierwszy, oparty na przeciwciałach zestaw do diagnozowania choroby SARS zostały stworzone przez członków projektu SEPSDA. Cztery spośród uczestniczących w projekcie laboratoriów to największe, czołowe, chińskie pracownie, które wniosły unikatowe próbki pochodzące od chińskich pacjentów na różnych etapach choroby. Serologiczne badania z pewnością doprowadzą do udoskonalenia metod diagnostycznych. Analiza genów koronawirusa na poziomie nukleotydowym oraz aminokwasowym za pomocą sekwencyjnych i zaawansowanych narzędzi bioinformatycznych oznaczy genetyczną zmienność wirusa. Bioinformatyczne badania wyselekcjonują również nowe białka wirusa, których zablokowanie może dać pozytywne efekty w leczeniu przeciwwirusowym. W projekcie SEPSDA postawiono sobie za cel określenie trójwymiarowej struktury wszystkich rozpuszczalnych białek lub całych ich domen. To strukturalno-genomowe podejście dostarczy podstaw do wirtualnego przesiewania dużych baz danych małych związków w celu odnalezienia substancji o potencjalnych właściwościach wpływania na funkcje białek wirusa lub białek komórek gospodarza, które z nimi współdziałają. Wykorzystane w realizacji projektu bazy danych zawierają między innymi struktury związków chemicznych używanych w tradycyjnej chińskiej medycynie. Wyselekcjonowani kandydaci na aktywne inhibitory zostaną przetestowani na liniach komórkowych, a następnie udoskonaleni chemicznie. Związki te po opatentowaniu zostaną zaoferowane utworzonej w przyszłości platformie przemysłowej zajmującej się chorobą SARS. Platforma ta powinna spowodować utworzeniem trwałej współpracy pomiędzy konsorcjum SEPSDA i przemysłem farmaceutycznym. Projekt realizuje następujące obszary priorytetowe: 1) nowe materiały, 2) zdrowie i życie, 3) technologie informacyjne.

**MicrobeArray (2005-2006)**
W ostatnich latach poznano sekwencje wielu genomów mikroorganizmów zaangażowanych w rozwój chorób o dużym znaczeniu dla zdrowia publicznego. Wraz z poznaniem całego zestawu genów danego organizmu możliwe stają się analizy wszystkich białek (proteomu) danego mikroorganizmu chorobotwórczego. W ostatniej dekadzie rozwinięto także technologie służące do wysokowydajnej produkcji znacznych ilości izolowanych i rekombinowanych białek. Dodatkowo wraz z rozwojem analiz z wykorzystaniem mikromacierzy (ang. microarray) powstały, zwalidowane już dziś klinicznie, metody służące do wykrywania obecności w surowicy przeciwciał przeciwko antygenom określonych patogenów. Odkrycia te dają możliwość badania naturalnej odpowiedzi immunologicznej przeciwko całym proteomom różnych mikroorganizmów. Połączenie informacji dotyczących genomu z narzędziami biologii molekularnej i immunologii pozwalają pomóc w identyfikacji tych antygenów, które mogą być użyte w diagnostyce serologicznej zakażeń i immunizacji (szczepienia ochronne). Podczas realizacji projektu MicrobeArray planowany jest rozwój narzędzi do identyfikacji mikrorganizmów o dużym znaczeniu medycznym – zarówno z powodu trudności w identyfikacji zakażeń przy użyciu tradycyjnych metod mikrobiologii lekarskiej, jak i znacznych różnic w postępowaniu terapeutycznym specyficznym dla danego zakażenia (M. pneumoniae, C. pneumoniae, L. pneumophila, coronavirus spp and P. falciparum). Antygeny tych groźnych patogenów otrzymane przy użyciu metod inżynierii genetycznej, jako rekombinowane białka, lub zestawy nakładających się syntetycznych peptydów, naniesione będą na mikromacierze. Surowica pobrana od osób z potwierdzonym przebytym zakażeniem danym patogenem będzie badana pod kątem interakcji z analizowanymi antygenami w celu identyfikacji nowych markerów diagnostycznych, czy wysoko immunogennych białek przydatnych przy wytwarzaniu szczepionek. Ten projekt znacząco zwiększy zasób własności intelektualnej małych i średnich przedsiębiorstw i przyczyni się do zwiększenia umiejętności i wiedzy jednostek badawczych. Planowany w czasie realizacji projektu rozwój technologii wysoko wydajnej ekspresji białek, oprogramowania analitycznego, syntezy białek na powierzchni mikromacierzy oraz czytników mikromacierzy pozwoli na stworzenie unikatowej integralnej platformy o wysokim potencjale zarówno komercyjnym, jak i naukowym. Dodatkowo wyniki projektu MicrobeArray mogą pomóc w rozszerzeniu naszej wiedzy na temat funkcjonowania humoralnej odpowiedzi immunologicznej człowieka,

jego interakcji z ogółem antygenów patogenu – przyczyniając się w ten sposób do wydajniejszego projektowania rekombinowanych szczepionek na podstawie dostępnych sekwencji genomowych.

**Diatomics (2005-2007)**

projektowania rekombinowanych szczepionek na podstawie dostępnych sekwencji genomowych.

Diatomics jest wieloośrodkowym projektem realizowanym przez kilkanaście laboratoriów naukowych i wdrożeniowych w ramach Szóstego Projektu Ramowego Komisji Europejskiej dla zrozumienia biologii okrzemek za pomocą metod genomiki funkcjonalnej.

Analizowane gatunki roślin jednokomórkowych należą do organizmów kosmopolitycznych i zamieszkują akweny morskie niemal całego globu. Jako model do analiz biologii okrzemek został wybrany gatunek Phaeodactylum tricornutum. Organizm ten posiada kilka cech czyniących go bardzo interesującym przedmiotem analiz, m.in. w odróżnieniu od innych okrzemek posiada możliwość rozmnażania płciowego oraz w stosunku do niego istnieje możliwość stosowania technik transformacji genetycznej. Są to główne powody, dlaczego zainicjowany został projekt poznania genomu tego organizmu, kierowany przez koordynatora projektu Diatomics - dr Chrisa Bowlera (Stazione Zoologica Anton Dohrn, Naples, Italy), a prowadzony w Joint Genome Institute, USA. Poznanie dokładnej biologii P.tricornutum będzie dodatkowo wspomagane przez informacje zebrane w czasie analizy genomu innego gatunku okrzemek - Thalassiosira pseudonana.

Projekt skupia się na wykorzystaniu informacji dotyczącej budowy genomów dwóch gatunków okrzemek morskich dla pełniejszego poznania funkcjonowania roślin. Organizmy te stanowią dużą część biomasy ekosystemów morskich i są odpowiedzialne za eliminacje znacznej części dwutlenku węgla z atmosfery Ziemi. Za podstawowe cele obrano dokładną charakteryzację procesów metabolicznych asymilacji dwutlenku węgla oraz makroelementów - zwłaszcza azotu i fosforu przez okrzemki. Dodatkowo projekt skupia się na analizie regulacji cyklu komórkowego okrzemek, próbując scharakteryzować mechanizmy odpowiedzialne za regulację występowania tzw. kwitnienia wód. Projekt obejmuje także dokładne studium ekspresji genów okrzemek w odpowiedzi na zmiany środowiska naturalnego.

Jako cel poboczny, w ramach projektu Diatomics planowana jest próba wytworzenia nowych transgenicznych odmian ryżu japońskiego, zawierających zidentyfikowane w ramach pierwszej części projektu geny korzystnie wpływające na sprawność fotosyntetyczną okrzemek.

Instytut Bioinfobank jest zaangażowany w kluczowy sposób w początkowej fazie projektu. Poprzez wykorzystanie publicznie dostępnych metod analizy sekwencyjnej oraz własnych technologii z dziedziny bioinformatyki (MetaBasic, Orfeus), mających na celu przygotowanie opisu funkcji poszczególnych białek z kręgu pakietów tematycznych projektu, umożliwione zostanie racjonalne zaplanowanie dalszych eksperymentów z użyciem technik biologii molekularnej i biologii komórki. Dokładna analiza funkcji białek zidentyfikowanych w genomach analizowanych organizmów pozwoli na postawienie hipotez roboczych odnośnie odmienności funkcjonowania poszczególnych procesów biologicznych w okrzemkach względem typowych modeli genetycznych organizmów roślinnych (rzodkiewnik - A.thaliana, ryż - O.sativa).

Dodatkowo w ramach projektu przewidziana jest systematyczne opracowanie zagadnienia transporterów błonowych substancji odżywczych do komórek analizowanych roślin. Problem ten wymaga opracowania metod identyfikacji specyficzności kanałów przezbłonowych, poprzez wykorzystanie metod analizy strukturalnej segmentów przezbłonowych i próbę określenia jakie właściwości musi spełniać jon lub cząsteczka, żeby przedostać się poprzez dany kanał.

Oprócz ww. analiz poszczególnych rodzin białkowych, w ramach projektu w dalszym ciągu będą rozwijane narzędzia porównywania rodzin białek, znajdujące zastosowanie w analizie struktury i funkcji. Dla potrzeb projektu dokonana zostanie walidacja baz danych sekwencji rodzin białek specyficznych dla roślin. Dotychczas rozwijane metody były optymalizowane dla sekwencji bakteryjnych (jako przykład białek głównie jednodomenowych) oraz sekwencji białek kręgowców (jako model białek wielodomenowych, typowych dla organizmów wyższych). Przystosowanie nowatorskiej metody wypracowanej w Instytucie Bioinfobank - narzędzia MetaBasic - do dokładnej analizy rodzin białek roślinnych, w oparciu o dostępne dane sekwencji białek badanych gatunków okrzemek, w znaczny sposób usprawni działanie tej metody i poszerzy możliwości jej stosowania. Dzięki temu możliwa będzie bardzo sprawną analizę genomów roślin, zwłaszcza kluczowych w produkcji żywności uprawnych, która może znaleźć późniejsze zastosowanie w projektach genomiki funkcjonalnej zsekwencjonowanych już genomów roślin wyższych (np.: ryż japoński), a także w przyszłych projektach tego typu (znacznie zaawansowany już dziś jest projekt badania genomu kukurydzy).

W ramach prac badawczych finansowanych w naszej instytucji ze środków własnych zrealizowane zostaną prace związane z dostosowaniem metod wykrywania odległych podobieństw między sekwencjami białek do

specyfiki organizmów podobnych do okrzemek. Ponadto przeprowadzone zostaną badania bioinformatyczne dotyczące potencjalnego wykorzystania genów scharakteryzowanych w badanych organizmach do modyfikacji roślin w celu zwiększenia wydajności hodowli.

**3. Wybrane publikacje za lata 2001-2004**

*Nazwa jednostki:* BioInfoBank Instytut

*Średnie zatrudnienie:* N=10 z wczytanych danych policzone jako: N = (5,00 + 6,00 + 13,00 + 16,00) / 4

**Wybrane czasopisma:**

***Publikacje w czasopiśmie wyróżnionym z listy filadelfijskiego Instytutu Informacji Naukowej:***
***(Autorzy z afiliacją Instytutu BioInfoBank są zaznaczeni na niebiesko)***

*1. Czasopismo:* Acta Biochimica Polonica
*Tytuł, rok, tom, str.:* Comparison of Proteins Based on the Segment Structural Similarity. 2004, 51(1), p. 161-172.
*Autor (autorzy):* Plewczynski D, Pas J, von Grotthuss M, Rychlewski L
*Punkty:* 10

*2. Czasopismo:* Bioinformatics
*Tytuł, rok, tom, str.:* The PDB-preview database: a repository of in-silico models of "on-hold" PDB entries. 2004 Apr 8
*Autor (autorzy):* Fischer D, Pas J, Rychlewski L
*Punkty:* 24

*3. Czasopismo:* Bioinformatics
*Tytuł, rok, tom, str.:* ToolShop: prerelease inspections for protein structure prediction servers. 2001 Dec;17(12):1240-1
*Autor (autorzy):* Rychlewski L.
*Punkty:* 24

*4. Czasopismo:* Bioinformatics
*Tytuł, rok, tom, str.:* Ligand-Info, searching for similar small compounds using index profiles. 2003 May 22;19(8):1041-2.
*Autor (autorzy):* von Grotthuss M, Pas J, Rychlewski L
*Punkty:* 24

*5. Czasopismo:* Bioinformatics
*Tytuł, rok, tom, str.:* 3D-Jury: a simple approach to improve protein structure predictions. 2003 May 22;19(8):1015-8.
*Autor (autorzy):* Ginalski K, Elofsson A, Fischer D, Rychlewski L
*Punkty:* 24

*6. Czasopismo:* BMC Bioinformatics
*Tytuł, rok, tom, str.:* RNA:(guanine-N2) methyltransferases RsmC/RsmD and their homologs revisited - bioinformatic analysis and prediction of the active site based on the uncharacterized Mj0882 protein structure. 2002 Apr 3;3(1):10.
*Autor (autorzy):* Bujnicki JM, Rychlewski L.
*Punkty:* 24

*7. Czasopismo:* Cell
*Tytuł, rok, tom, str.:* mRNA cap-1 methyltransferase in the SARS genome. 2003 Jun 13;113(6):701-2.
*Autor (autorzy):* von Grotthuss M, Wyrwicz LS, Rychlewski L
*Punkty:* 24

*8. Czasopismo:* Comb Chem High Throughput Screen
*Tytuł, rok, tom, str.:* Ligand.Info small-molecule Meta-Database. 2004 Dec;7(8):757-61.
*Autor (autorzy):* von Grotthuss M, Koczyk G, Pas J, Wyrwicz LS, Rychlewski L
*Punkty:* 24

*9. Czasopismo:* DNA Repair (Amst)
*Tytuł, rok, tom, str.:* Fold-recognition analysis predicts that the Tag protein family shares a

common domain with the helix-hairpin-helix DNA glycosylases. 2002 May 30;1(5):391-5.
*Autor (autorzy):* Bujnicki JM, Rychlewski L.
*Punkty:* 24

*10. Czasopismo:* FEBS Lett
*Tytuł, rok, tom, str.:* Structure prediction, evolution and ligand interaction of CHASE domain. 2004 Oct 22;576(3):287-90.
*Autor (autorzy):* Pas J, von Grotthuss M, Wyrwicz LS, Rychlewski L, Barciszewski J
*Punkty:* 24

*11. Czasopismo:* Gene
*Tytuł, rok, tom, str.:* Molecular phylogenetics of the RrmJ/fibrillarin superfamily of ribose 2'-O-methyltransferases. 2003 Jan 2;302(1-2):129-38.
*Autor (autorzy):* Feder M, Pas J, Wyrwicz LS, Bujnicki JM
*Punkty:* 20

*12. Czasopismo:* Nucleic Acids Res
*Tytuł, rok, tom, str.:* Detecting distant homology with Meta-BASIC. 2004, 32, W576-W581.
*Autor (autorzy):* Ginalski K, von Grotthuss M, Grishin NV, Rychlewski, L
*Punkty:* 24

*13. Czasopismo:* Nucleic Acids Res
*Tytuł, rok, tom, str.:* Detection of reliable and unexpected protein fold predictions using 3D-Jury. 2003 Jul 1;31(13):3291-2.
*Autor (autorzy):* Ginalski K, Rychlewski L
*Punkty:* 24

*14. Czasopismo:* Nucleic Acids Res
*Tytuł, rok, tom, str.:* ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. 2003 Jul 1;31(13):3804-7.
*Autor (autorzy):* Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L
*Punkty:* 24

*15. Czasopismo:* Protein Eng
*Tytuł, rok, tom, str.:* In silico identification, structure prediction and phylogenetic analysis of the 2'-O-ribose (cap 1) methyltransferase domain in the large structural protein of ssRNA negative-strand viruses 2002 Feb;15(2):101-8.
*Autor (autorzy):* Bujnicki JM, Rychlewski L.
*Punkty:* 20

*16. Czasopismo:* Protein Eng
*Tytuł, rok, tom, str.:* The 2002 olympic games of protein structure prediction. 2003 Mar;16(3):157-60.
*Autor (autorzy):* Fischer D, Rychlewski L
*Punkty:* 20

*17. Czasopismo:* Proteins
*Tytuł, rok, tom, str.:* Application of 3D-Jury, GRDB, and Verify3D in fold recognition. 2003 V53, S6:418-423.
*Autor (autorzy):* von Grotthuss M, Pas J, Wyrwicz L, Ginalski K, Rychlewski L
*Punkty:* 24

*18. Czasopismo:* Proteins
*Tytuł, rok, tom, str.:* Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. 2003 V53,S6:410-417
*Autor (autorzy):* Ginalski K, Rychlewski L
*Punkty:* 24

*19. Czasopismo:* Science
*Tytuł, rok, tom, str.:* Predicting Protein Structures Accurately. 2004 Jun 11;304:1596.
*Autor (autorzy):* von Grotthuss M, Wyrwicz LS, Pas J, Rychlewski L

*Punkty:* 24

*20. Czasopismo:* Science
*Tytuł, rok, tom, str.:* How unique is the rice transcriptome? 2004 Jan 9;303:168.
*Autor (autorzy):* Wyrwicz LS, von Grotthuss M, Pas J, Rychlewski L
*Punkty:* 24

**4. Pozytywne aspekty kategoryzacji Instytutu BioInfoBank**

Kategoryzacja jednostki „Instytut BioInfoBank" będzie miało między innymi następujące pozytywne rezultaty:

**a) Zahamowanie emigracji wysoko wykształconej kadry naukowej.**
W skład osób inwestujących wiele starań w stworzenie przedstawianej jednostki bioinformatycznej w Polsce wchodzi szereg młodych polskich naukowców. Często są to studenci którzy dzięki nam decyzują się na karierę zawodową. Wiele starszych pracowników z powodzeniem mogłoby znaleźć dogodne warunki pracy za granicą (n.p. autor projektu dr Leszek Rychlewski, dr Krzysztof Ginalski oraz dr Dariusz Plewczyński otrzymali oferty objęcia atrakcyjnych i niezależnych pozycji profesorskich na uczelniach w USA). Finansowanie przez Instytut oraz dostęp do technologii oferowanej przez Instytut jest jednym z istotnych czynników wpływających na decyzję o pobycie w Polsce. Stworzenie nowoczesnej jednostki naukowej w Polsce, na której działalność będą mieli istotny wpływ i która będzie w stanie przedstawić im atrakcyjny, perspektywiczny plan ich dalszej kariery, nakłania młodą kadrę do pozostania w kraju i przyczyni się do rozwoju technologii w Polsce. Nowa jednostka będzie również w stanie ściągać polskich naukowców do ojczyzny.

**b) Stworzenie pilotażowej jednostki naukowej nowego typu.**
Jednostki badawcze działające wewnątrz przedsiębiorstw komercyjnych mają utrudniony dostęp do najnowszych osiągnięć naukowych, gdyż muszą się często liczyć z istotnie zwiększonymi opłatami licencyjnymi w porównaniu do jednostek „non for profit". Jednostki naukowe z drugiej strony, posiadają dużo wygodniejszy dostęp do narzędzi badawczych, jednak cechują się relatywnie niską efektywnością starań o wdrożenie rezultatów swoich badań. Krzyżówka obu typów tych jednostek badawczych, prywatna jednostka „non for profit", jest idealnym rozwiązaniem usprawniającym efektywność i użyteczność badań, które sprawdziło się w wielu rozwiniętych krajach przemysłowych.

**c) Zwiększenie konkurencyjności w nauce polskiej.**
Do krajobrazu instytucji działających na rzecz nauki dojdzie nowa, sprawna jednostka, cechująca się nowoczesnym i efektywnym systemem administracji i zarządzania oraz silnie motywacyjnym systemem wynagradzania współpracowników. Jednostka taka będzie zabiegać o fundusze z Ministerstwa Nauki i Informatyzacji i konkurować z innymi renomowanymi polskimi jednostkami. Jeżeli projekt stworzenia jednostki badawczej nowego typu okaże się sukcesem, będzie to dobrym sygnałem i inspiracją do udoskonalania aktualnej infrastruktury administracji nauki w Polsce. Instytut BioInfoBank z powodu specyfiki nauk bioinformatycznych (ich relatywnie niskiego kosztu przy wysokiej wydajności) jest idealnym projektem pilotażowym do tego celu.

**d) Zwiększenie renomy nauki polskiej.**
Jednostka Instytut BioInfoBank zdobyła już pewną zauważalną renomę w branży bioinformatycznej na świecie. Jednostka ta jest źródłem znaczącej liczby publikacji w bardzo znanych czasopismach naukowych, prowadzi renomowane międzynarodowe konkursy badawcze, uczestniczy z powodzeniem w sprawdzianach umiejętności w dziedzinie przewidywania struktury białek, udostępnia unikalne i popularne narzędzia bioinformatyczne oraz prowadzi współpracę naukową na szeroką światową skalę, również w ramach wielu projektów europejskich. Rezultatami naszej działalności mogą być stworzenie nowej polskiej marki i likwidacja negatywnych stereotypów dotyczących zacofania polskiego przemysłu.