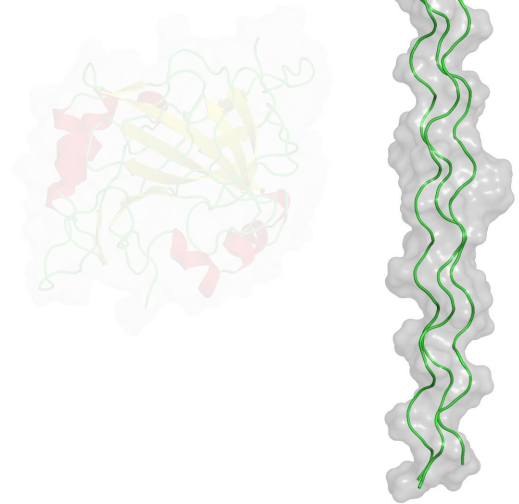


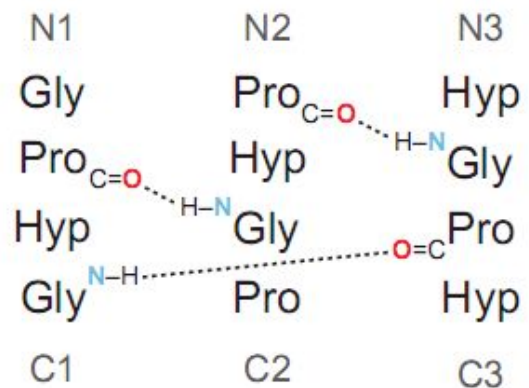
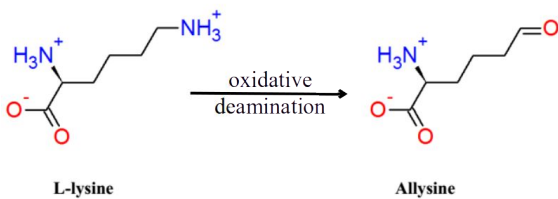
Strukturalne rodzaje białek

- **Globularne**
białko globularne, białko o kształcie kulistym lub sferoidalnym, zbudowane z gęsto pofalowanych lub pozwijanych łańcuchów peptydowych; wartość stosunku osiowego cząsteczki nie przekracza 3 : 1 lub 4 : 1; do b. g. należy ogromna większość białek, m.in. albuminy i globuliny oraz wszystkie enzymy.
- **Fibrylarne**
białka proste o strukturze włóknkowej stanowiące podstawowy materiał budulcowy organizmów zwierzęcych. Są to typowe białka o budowie włóknistej, dzięki temu pełnią funkcje podporowe. Do tej grupy białek należy keratyna, kolagen, miozyna i fibroina.



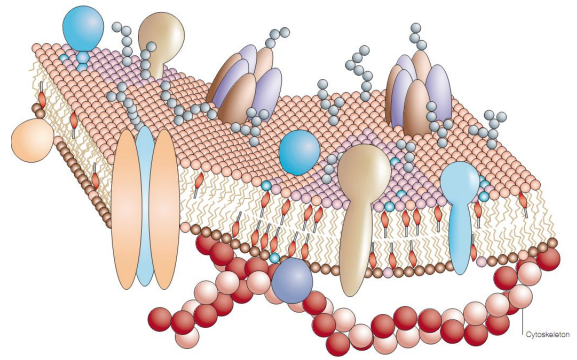
Kolagen

- **Stabilność:**
 - wiązania wodorowe
 - cross-link związany ze starzeniem



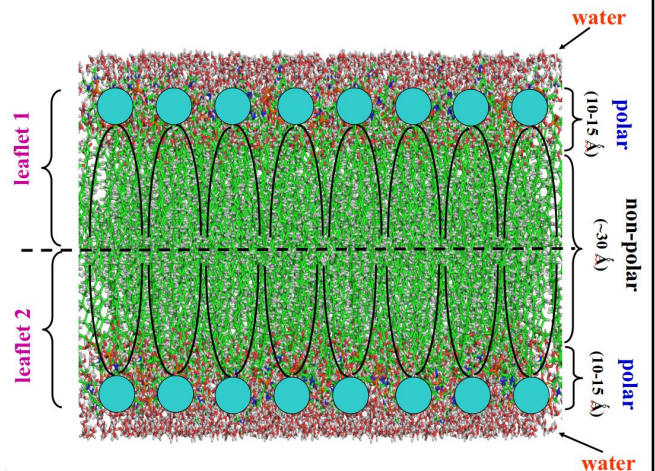
Białka błonowe

- Białka błonowe to powszechne białka, które są częścią błon biologicznych lub wchodzą z nimi w interakcje.
- Białka błonowe dzielą się na kilka szerokich kategorii w zależności od ich lokalizacji.
 - Integralne białka błonowe są stałą częścią błony komórkowej i mogą przenikać przez błonę (transbłonowe) lub wiązać się z jedną lub drugą stroną błony (integralne monotopowe).
 - Peryferyjne białka błonowe są przejściowo związane z błoną komórkową.



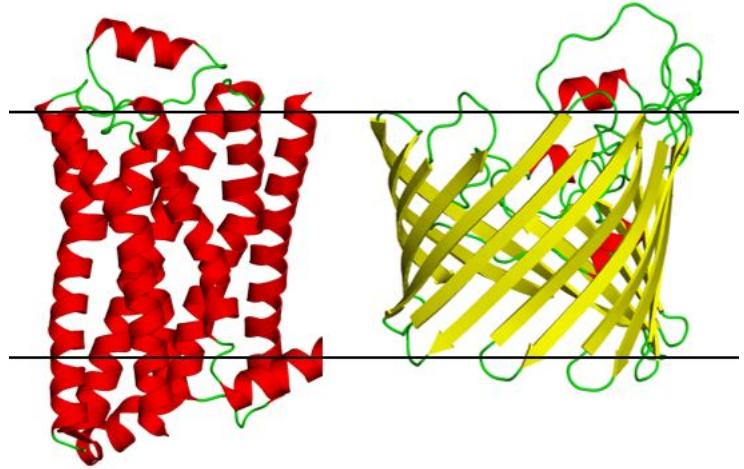
Błona komórkowa

- Błona biologiczna, biomembrana lub błona komórkowa to selektywnie przepuszczalna błona, która oddziela wnętrze komórki od środowiska zewnętrznego lub tworzy przedziały wewnątrzkomórkowe, służąc jako granica między jedną częścią komórki a drugą.
- Błony biologiczne, w postaci błon komórek eukariotycznych, składają się z dwuwarstwy fosfolipidowej z wbudowanymi, integralnymi i peryferyjnymi białkami wykorzystywanymi w komunikacji i transporcie substancji chemicznych i jonów. Większość lipidów w błonie komórkowej zapewnia płynną matrycę dla białek, które mogą się obracać i dyfundować na boki w celu fizjologicznego funkcjonowania.
- Białka są przystosowane do wysokiej płynności środowiska dwuwarstwy lipidowej z obecnością pierścieniowej otoczki lipidowej, składającej się z cząsteczek lipidów ściśle związanych z powierzchnią integralnych białek błonowych. Błony komórkowe różnią się od izolujących tkanek utworzonych przez warstwy komórek, takich jak błony śluzowe, błony podstawne i błony surowicze.



Białka błonowe struktura

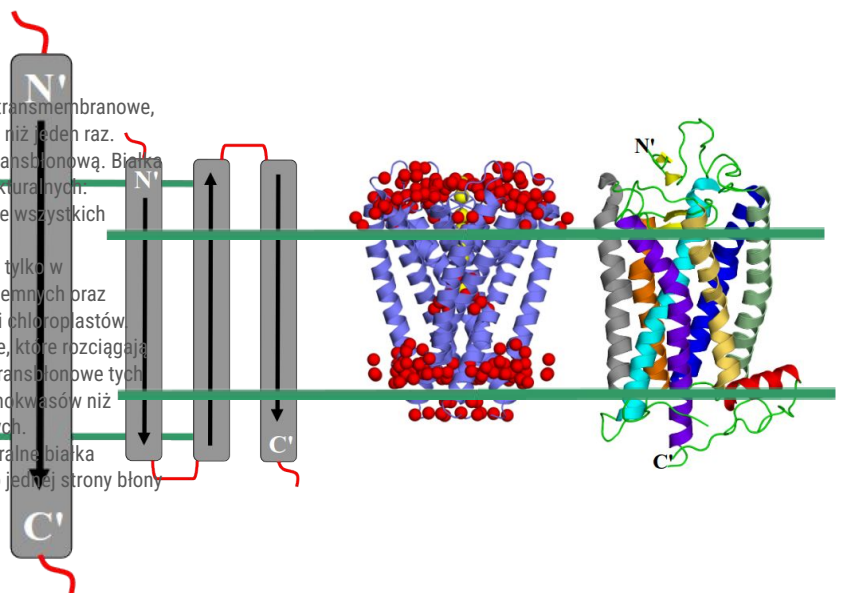
- >90 %: α helikalne



Białka błonowe struktura

Bitopic

- >90 %: α helikalne
- Integralne białka politopowe to białka transmembranowe, które rozciągają się przez błonę więcej niż jeden raz. Białka te mogą mieć różną topologię transbłonową. Białka te mają jedną z dwóch architektur strukturalnych.
- Białka wiązki helisy, które są obecne we wszystkich typach błon biologicznych;
- Białka beta-beczkowe, które występują tylko w zewnętrznych błonach bakterii Gram-ujemnych oraz zewnętrznych błonach mitochondriów i chloroplastów
- Białka bitopowe to białka transbłonowe, które rozciągają się w poprzek błony tylko raz. Helikty transbłonowe tych białek mają znacznie inny rozkład aminokwasów niż helikty transbłonowe białek politopowych.
- Integralne białka monotopowe to integralne białka błonowe, które są przyłączone tylko do jednej strony błony i nie rozciągają się na całej długości.



Keratyna

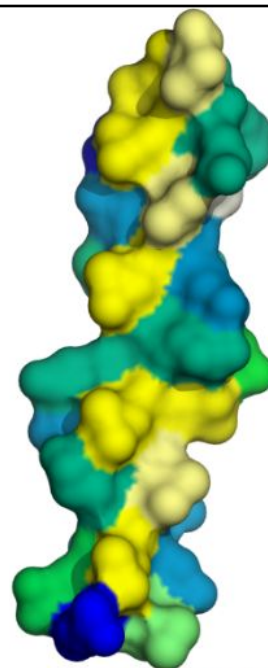
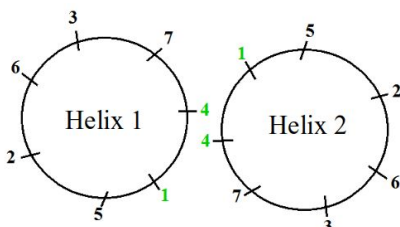
- Powtarzalny motyw: HxxHCxC

H-hydrofobowe

C-naładowane

P-polarne

a	b	c	d	e	f	g
H	P	P	H	C	P	C



L I F V A G C S T W P Y Q H K N E D R
 Hydrophobic Hydrophilic

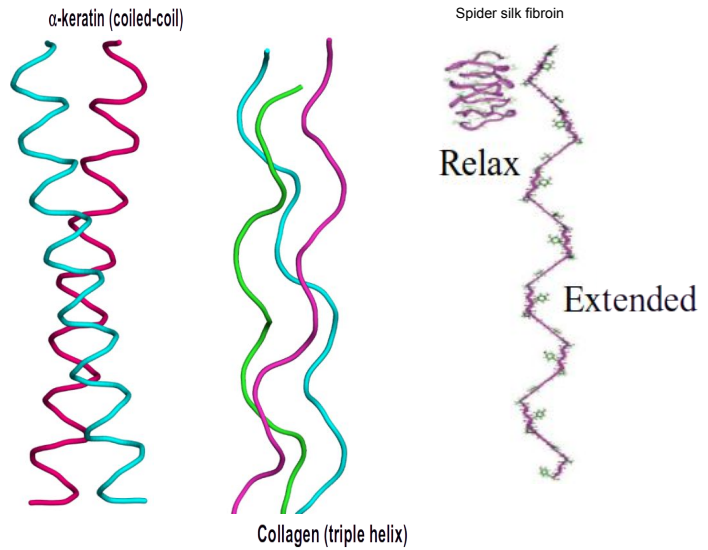
Kolagen

- Kolagen (/ 'kolədʒən/) jest głównym białkiem strukturalnym w macierzy zewnątrzkomórkowej występującej w różnych tkankach łącznych organizmu.
- Jako główny składnik tkanki łącznej, jest najobficiej występującym białkiem u ssaków, stanowiąc od 25% do 35% zawartości białka w całym organizmie.
- Kolagen składa się z aminokwasów połączonych ze sobą, tworząc potrójną helisę wydłużonych włókien znaną jako helisa kolagenowa. Występuje głównie w tkance łącznej, takiej jak chrząstka, kości, ścięgna, więzadła i skóra.
- Witamina C jest niezbędna do syntezy kolagenu, a witamina E zwiększa jego produkcję.

Typ	Budowa	Występowanie	Inne cechy
Kolageny fibylarne (włókniste)			
I	[$\alpha 1(I)$] ₂ [$\alpha 2(I)$]	skóra, ścięgna, więzadła, kości, zębina, inne tkanki łączne niebędące chrząstką	włókno o dl. 300 nm
II	[$\alpha 1(II)$] ₃	chrząstki, ciało szkliste, jądro miążdżyste (w krążku międzykręgowym)	włókno o dl. 300 nm
III	[$\alpha 1(III)$] ₃	rozciągliwe tkanki łączne jak skóra, mięśnie, układ naczyniowy; często razem z typem I	włókno o dl. 300 nm
V	[$\alpha 1(V)$] ₂ [$\alpha 2(V)$], [$\alpha 1(V)$] ₃	rogówka, zęby, kości, łożysko, skóra, mięśnie gładkie; często razem z typem I	włókno o dl. 390 nm, N-końcowy fragment globularny
Kolageny tworzące układy sieciowe			
IV	[$\alpha 1(IV)$] ₂ [$\alpha 2(IV)$]	wszystkie błony podstawne	
VIII	[$\alpha 1(VIII)$] ₃ , [$\alpha 2(VIII)$] ₃ , [$\alpha 1(VIII)$] ₂ [$\alpha 2(VIII)$] ³⁵⁾	śródbłonek, błona Descemeta	tworzy heksagonalne układy sieciowe
X	[$\alpha 1(X)$] ₃	chrząstka hipertroficzna	tworzy heksagonalne układy sieciowe
Kolageny kotwiczące			
VII	[$\alpha 1(VII)$] ₃	pódesmosomy	najdłuższa domena trójhelikalna o dl. 420 nm
Kolageny towarzyszące włóknom o przerywanej strukturze helisowej (FACIT)			
IX	[$\alpha 1(IX)$] ₂ [$\alpha 2(IX)$] ₃ [$\alpha 3(IX)$]	chrząstki, ciało szkliste; towarzyszy kolagenowi typu II	3 krótkie domeny kolagenowe zawierające 4 regiony globularne ¹¹⁾
XII	$\alpha 1(XII)$ ₃ ¹¹⁾	towarzyszy kolagenowi typu I	
Kolageny transblonowe			
XIII	[$\alpha 1(XIII)$] ₃	wiele tkanek, m.in. płytki nerwowo-mięśniowa	kolageny transblonowe mają N-końiec zatopiony wewnątrz komórki i domenę hydrofobową przechodzącą przez dwuwarstwą lipidową błony ¹¹⁾
XVII	[$\alpha 1(XVII)$] ₃	nabłonek, hemidesmosomy	
Multipleksyny – kolageny o licznych nieciągłych domenach o budowie helikalnej			
XV	[$\alpha 1(XV)$] ₃	występuje w proteoglikanach, gdzie stanowi rdzeń białkowy połączony z siarczanem chondroilny; towarzyszy kolagenom błony podstawnej	stabilizuje mięśnie szkieletowe i mikronaczynia
XVIII	[$\alpha 1(XVIII)$] ₃	towarzyszy kolagenom błony podstawnej	utrzymuje strukturalną integralność błony podstawnej; bliski homolog kolagenu typu XV

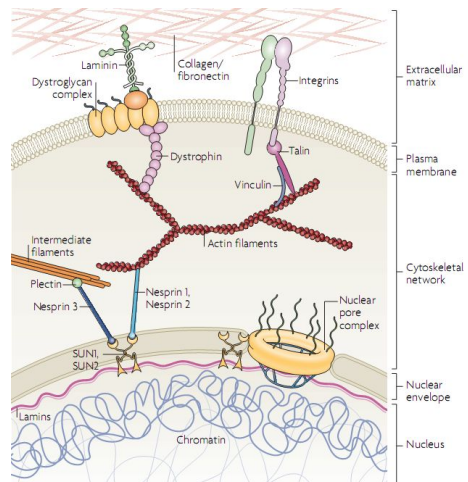
Właściwości ogólne

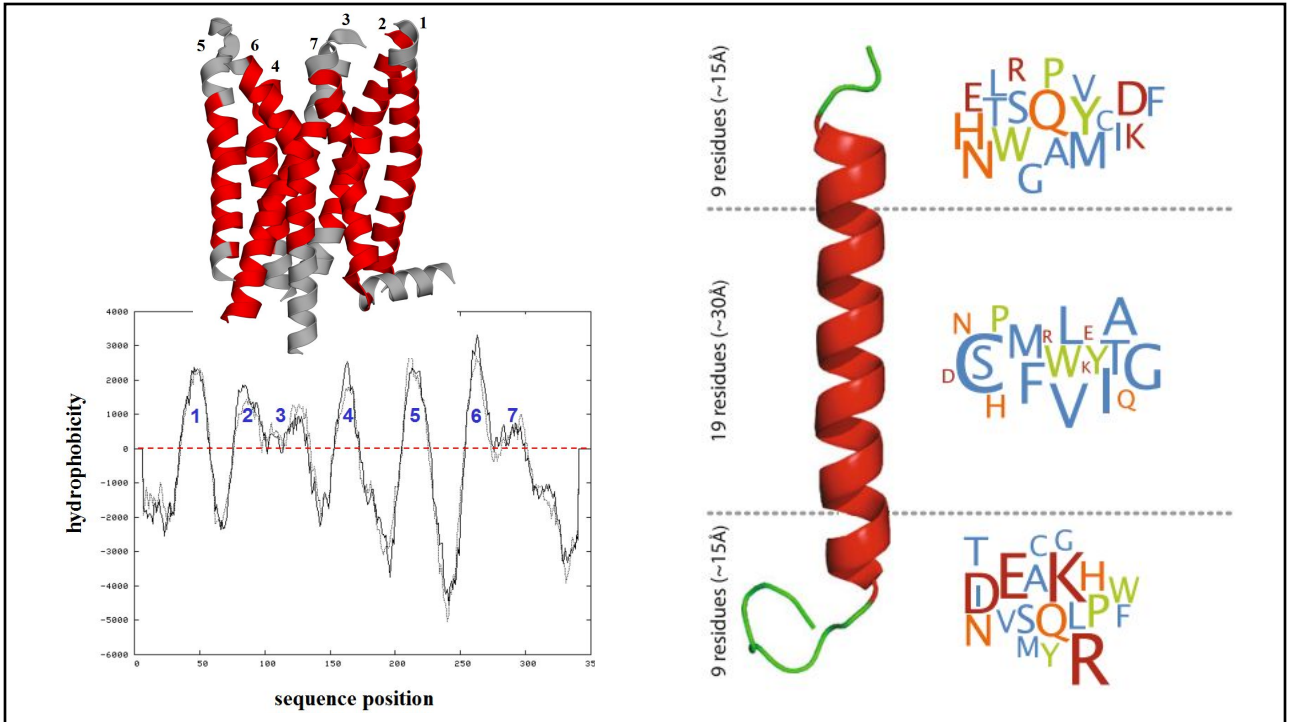
- Superstruktura drugorzędowa
 - nadrzędne elementy α lub β
- Hydrofobowość
- Zdolność tworzenia makrostruktur czwartorzędowych



Wykorzystanie białek fibrylarnych

- Wsparcie mechaniczne i ochrona
 - cytoszkielet, np. keratyna
 - macierz pozakomórkowa, np. kolagen
- Budowa struktur komórkowych
- Komunikacja wewnątrzkomórkowa
- 'Ruch'
- Budowa struktur zewnętrznych, np. jedwab
- Funkcje regulacyjne
 - pigmentacja
 - wzrost włosów
 - regeneracja nabłonka





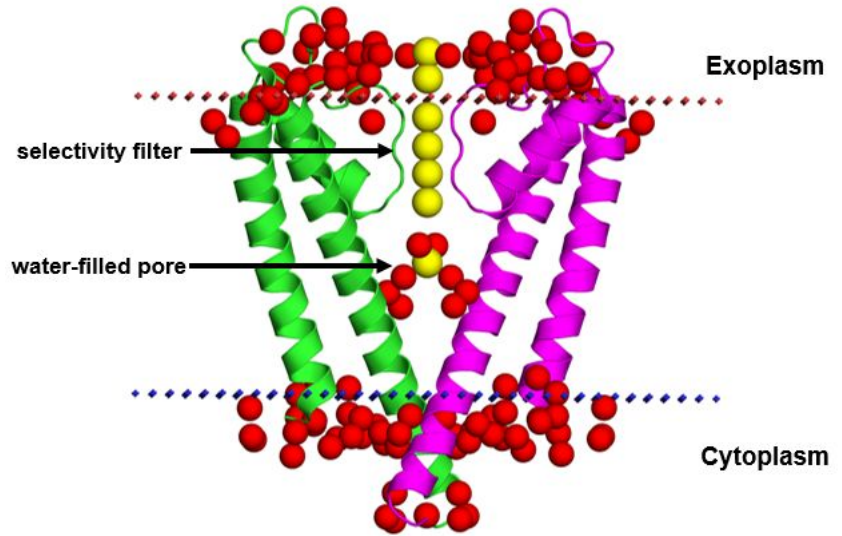
Dynamika

- Zmiana konformacji związana z funkcją:
 - kanał/transporter musi wykazywać selektywność
 - receptor musi posiadać formę aktywną i nieaktywną

The figure shows a 3D ribbon diagram of a protein structure. The protein is primarily red, with a central region highlighted in blue. A specific residue, labeled "Pro", is highlighted in green. An inset diagram shows a close-up of the Proline residue (Pro-14) in a red ribbon structure, with a green circle highlighting the residue. The inset is positioned between two horizontal green bars.

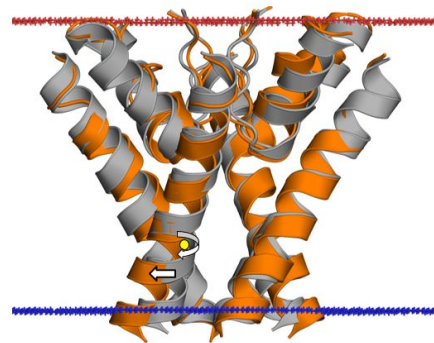
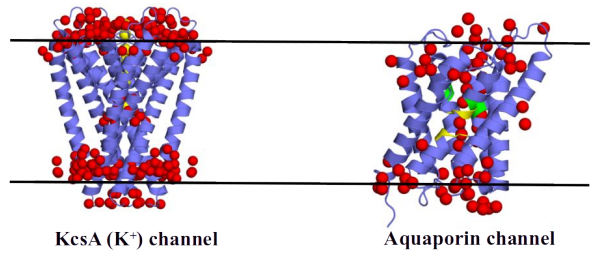
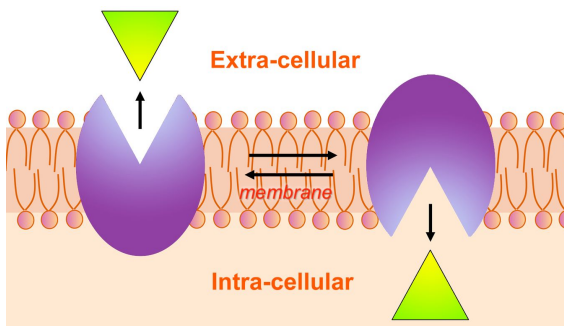
Białka transportowe

- Selektywność



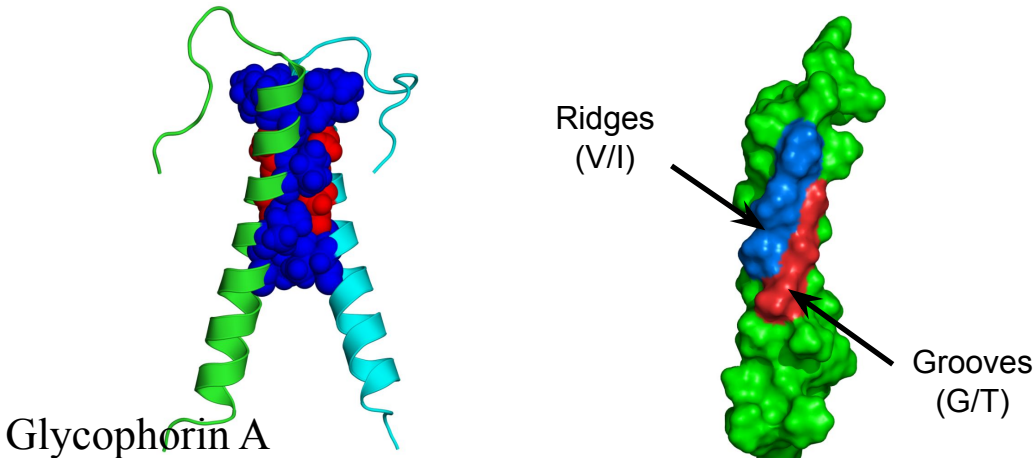
Białka transportowe

- Reaktywność



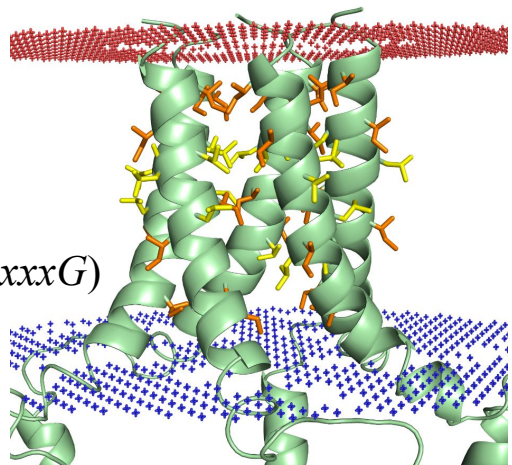
Efektywność upakowania w błonie

- **Grzebiety (V/L/I) & doliny (G/S/T/A)** [motyw **GxxG**]
- **Nachylenie (~20°)**



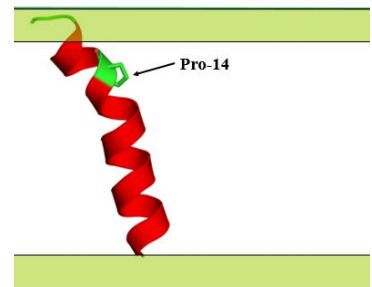
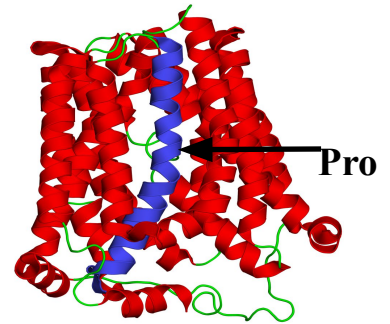
Efektywność upakowania w błonie

- **Inne motywy stabilizujące**
 - Suwaki leucynowe/izoleucynowe
 - Suwaki serynowe
 - Suwaki glicynowe (motyw *GxxxGxxxG*)



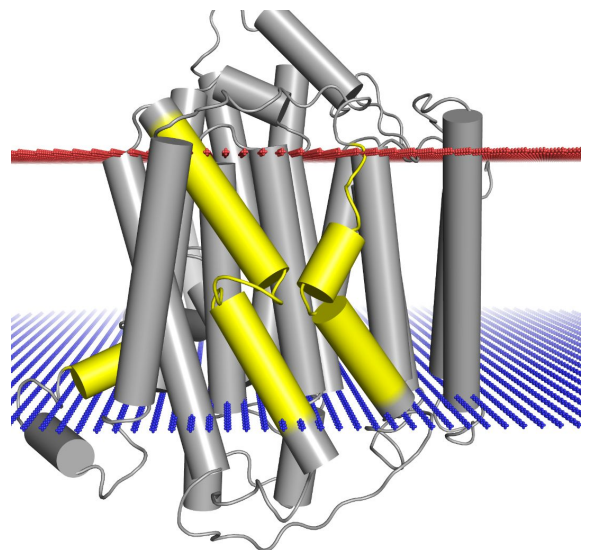
Elementary dynamiczne

- **Gly, Pro** znajdujące się w załamaniach
- **Pro** w elementach transbłonowych (TM) służą:
 - rodzaj zawiasów
 - odpowiadają za aktywację funkcji transportowych
 - zmianę konformacji receptorów



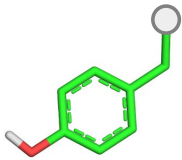
Elementary dynamiczne

- **Nieciągłość helis (około 40%):**
 - helisy 3_{10} ,
 - π -helices,
 - przerwy.
- **Konsekwencje dudowy mechanicznej:**
 - optymalizacja odległości
 - tworzenie miejsc aktywnych

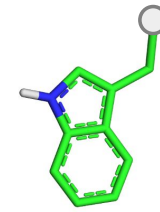
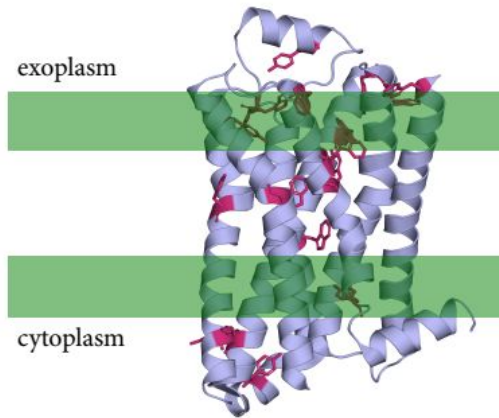


Skład jakościowy

- **Pas aromatyyczny - zakotwiczenie**



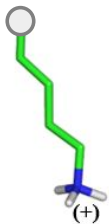
Tyrosine
(Tyr, Y)



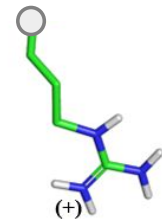
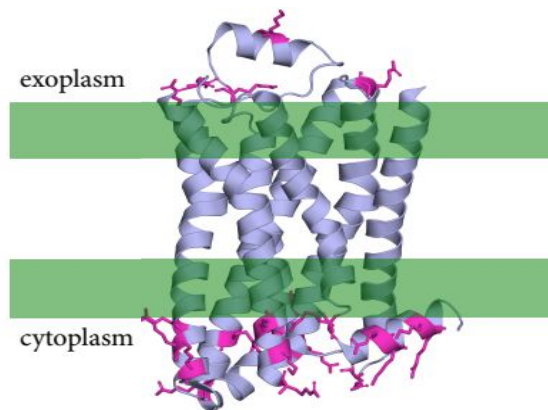
Tryptophan
(Trp, W)

Skład jakościowy

- **Arg, Lys – oddziaływania zewnętrzne**



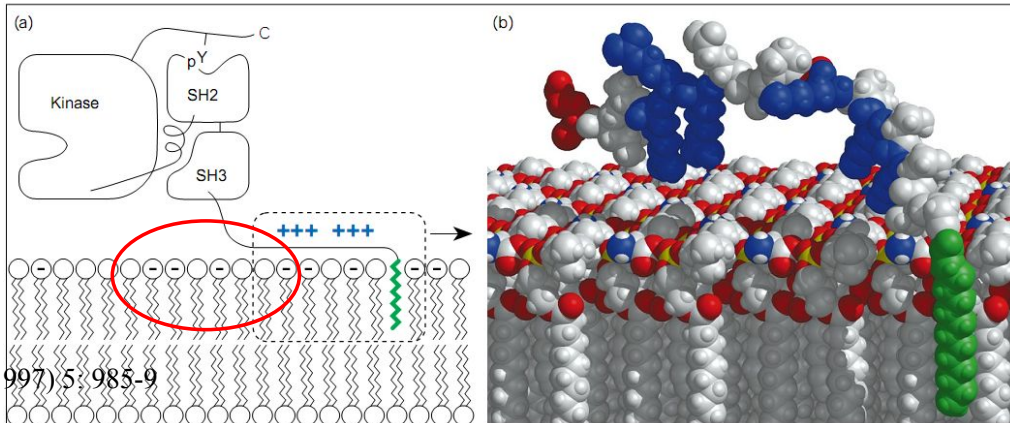
Lysine
(Lys, K)



Arginine
(Arg, R)

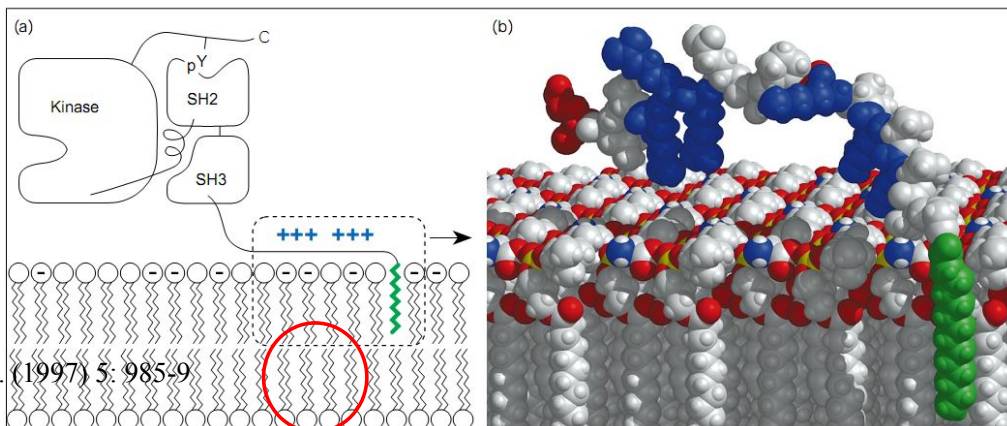
Białka
związane z błonami

- oddziaływania elektrostatyczne



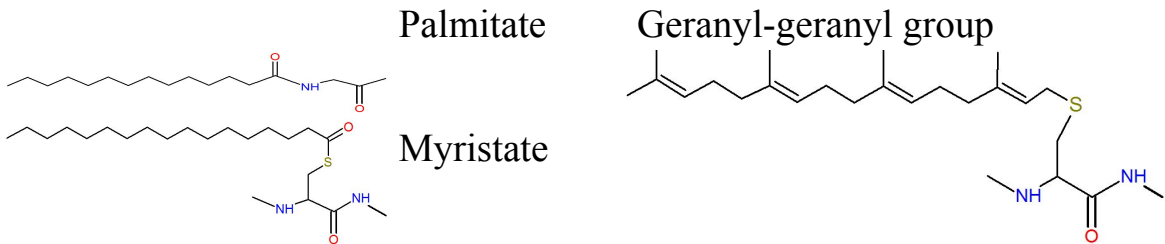
Białka
związane z błonami

- oddziaływania kowalencyjne



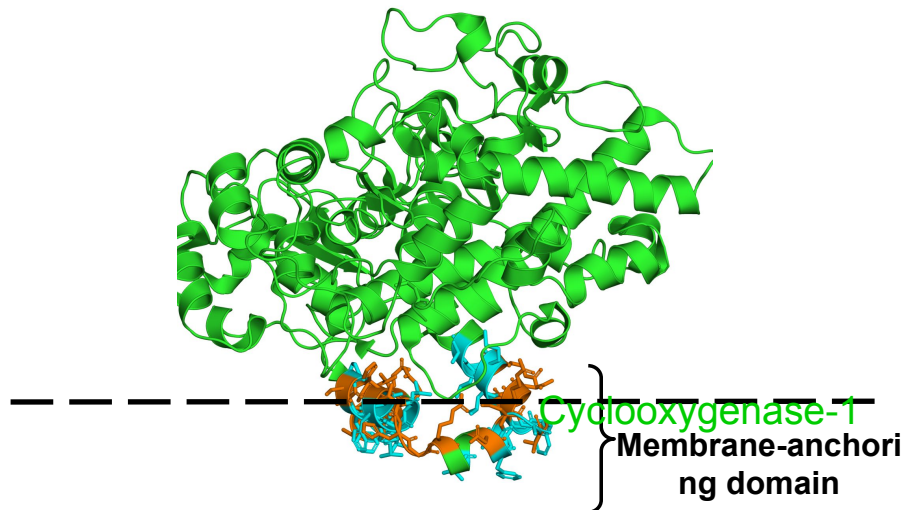
Białka
związane z błonami

- **oddziaływania kowalencyjne:**
- kwas tłuszczowy związany z N-terminalną glicyną lub cysteiną/seryną
- hydrofobowej reszty farnezylowej lub geranylogeranylowej przyłączona do cysteiny



Kotwiczenie
za pomocą domeny

- **Nonpolar**

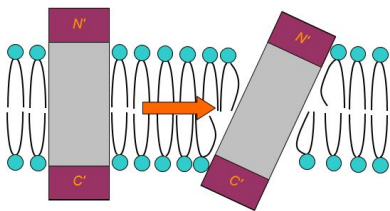


Interakcje białko-błona

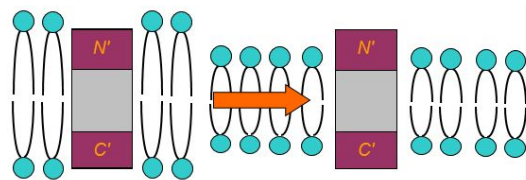
Wpływ błony na białko

- **Niedopasowanie hydrofobowe**

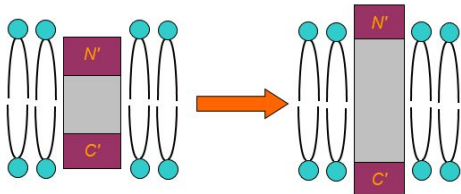
Skrzywienie



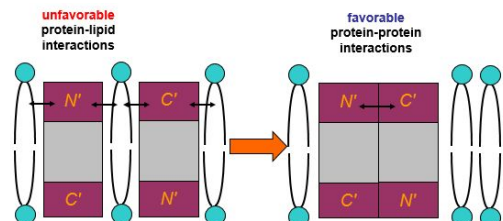
Dyfuzja pozioma



Zmiana konformacji

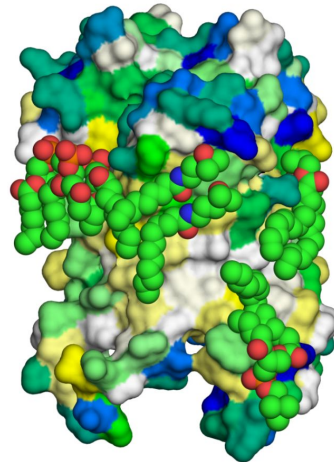


Oligomeryzacja



Wpływ błony na białko

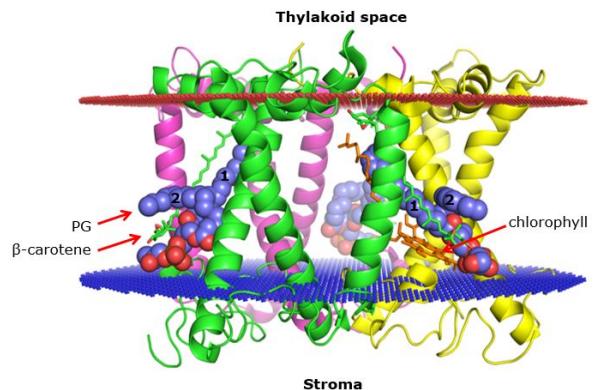
- **Pojedyncze komponenty błony mogą oddziaływać silnie z białkiem (np. artefakty krystalizacyjne).**



L I F V A G C S T W P Y Q H K N E D R
Hydrophobic Hydrophilic

Wpływ kw. tłuszczowych na białko

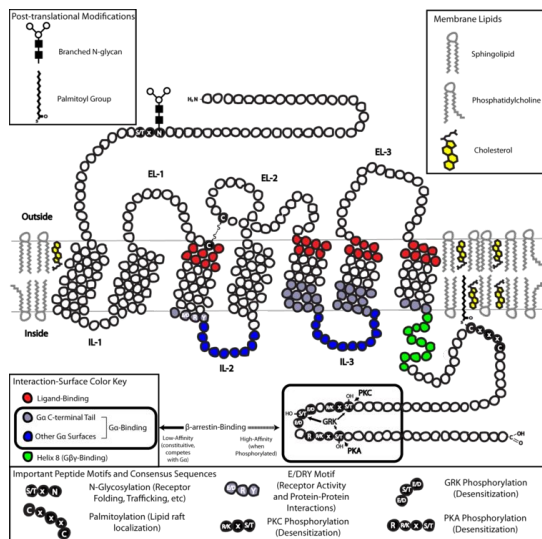
- **Interakcje pomiędzy kwasami tłuszczowymi a białkami błonowymi są niezbędne dla utrzymania:**
 - **stabilności**
 - **poprawności**
 - **fałdowania**
 - **aktywności biologicznej**



Receptory sprzężone z białkami G

- GPCRs (ang. G protein coupled receptors) stanowią najliczniejszą i bardzo zróżnicowaną grupę białek błonowych odpowiedzialnych za przekazywanie sygnałów przez dwuwarstwą lipidową do miejsc efektorowych znajdujących się wewnątrz komórki.
- Sekwencjonowanie ludzkiego genomu ujawniło występowanie ok. 800 różnych typów receptorów należących do rodziny GPCR (geny kodujące receptory GPCR stanowią powyżej 3% ludzkiego genomu), a ponad połowa z nich wykazuje potencjalne znaczenie dla przemysłu farmaceutycznego.
- Analiza porównawcza sekwencji receptorów GPCR oraz badania funkcji poszczególnych typów receptorów doprowadziły do podziału tej rodziny białek na klasy, które oznaczono literami od A do F.
- Pierwsza, najliczniejsza klasa A (nazywana również klasą podobnych do rodopsyny – ang. rhodopsin like), obejmująca ponad 80% wszystkich GPCR, to klasa receptorów rodopsyno-podobnych. Receptory wchodzące w skład tej grupy są jednymi z najlepiej zbadanych. Zaliczamy do nich, obok rodopsyny, między innymi: receptory adrenergiczne, opioidowe, adenozynowe, kanabinoidowe, receptory chemokin, dopaminowe i histaminowe.
- Klasę B stanowią receptory sekretyno-podobne. Do klasy C zaliczamy receptory glutaminergiczne i feromonowe.
- Kolejne klasy D i E tworzą odpowiednio receptory feromonów grzybów oraz receptory cAMP.
- Ostatnia klasa F to receptory frizzled/smoothened. Klasyfikacja ta pokrywa się w większości z nową klasyfikacją GRAFS, opartą na badaniach filogenetycznych. Nazwa GRAFS pochodzi od pierwszych liter wyodrębnionych rodzin receptorów, do których należą odpowiednio: receptory Glutaminergiczne, Rodopsyno-podobne, Adhezyjne, Frizzled i smakowe oraz Sekretyno-podobne.

GPCRs



The Nobel Prize in Chemistry 2012



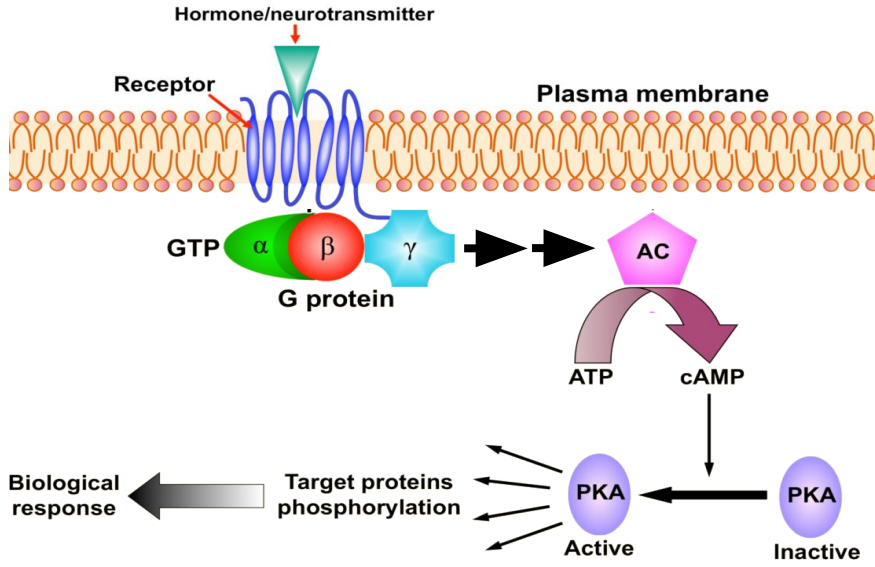
© The Nobel Foundation, Photo: U. Montan
Robert J. Lefkowitz
 Prize share: 1/2



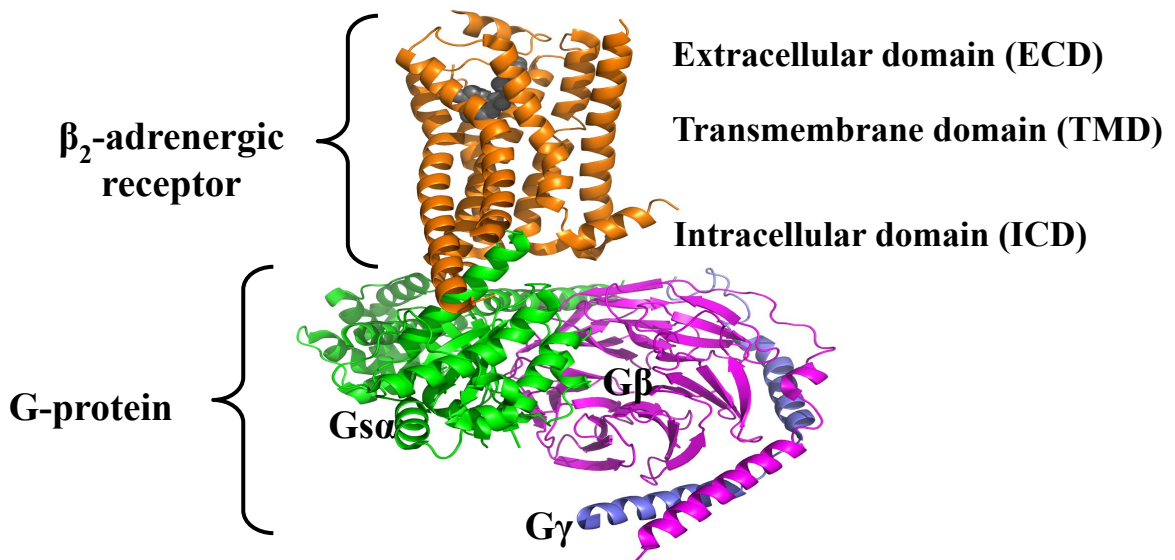
© The Nobel Foundation, Photo: U. Montan
Brian K. Kobilka
 Prize share: 1/2

The Nobel Prize in Chemistry 2012 was awarded jointly to Robert J. Lefkowitz and Brian K. Kobilka "for studies of G-protein-coupled receptors"

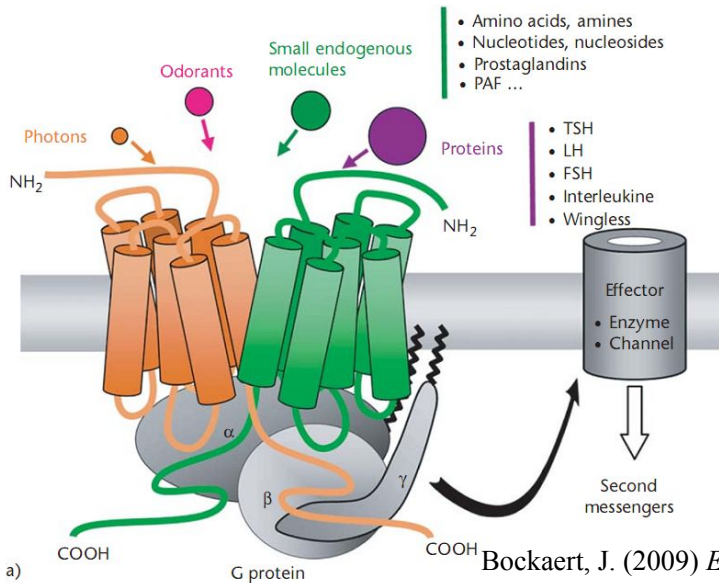
GPCRs ścieżka sygnalizacyjna



The GPCR-G protein complex



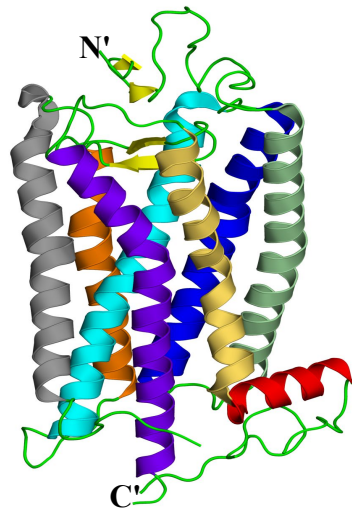
GPCRs ligandy i efektery



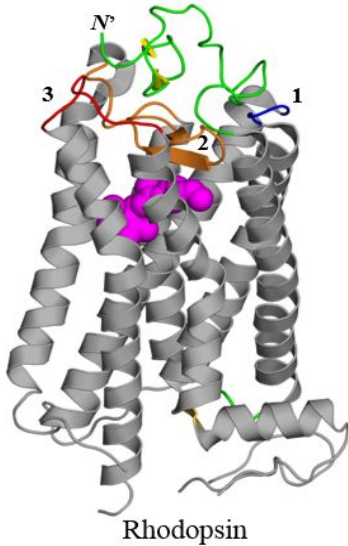
Bockaert, J. (2009) *ELS*. John Wiley & Sons, Ltd

GPCRs podobieństwo strukturalne

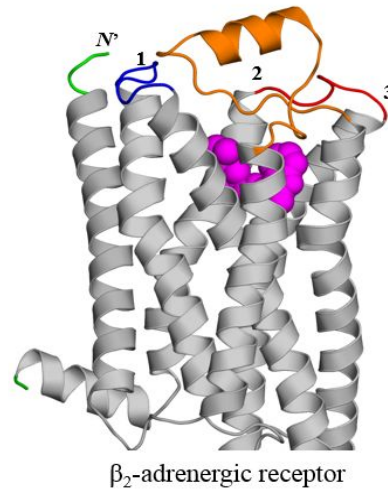
- **GPCRs mają podobne struktury pomimo niskiej zgodności sekwencji**
 - Rhodopsin & β_2 -AR: 21% zgodności,
 - TM r.m.s.d = 1.6 Å
- **TMD jest najlepiej zachowaną częścią z 7 TM helisami**



GPCRs różnice strukturalne



(ECL2 is long and blocks entrance)



(ECL2 is a short helix, entrance is open)

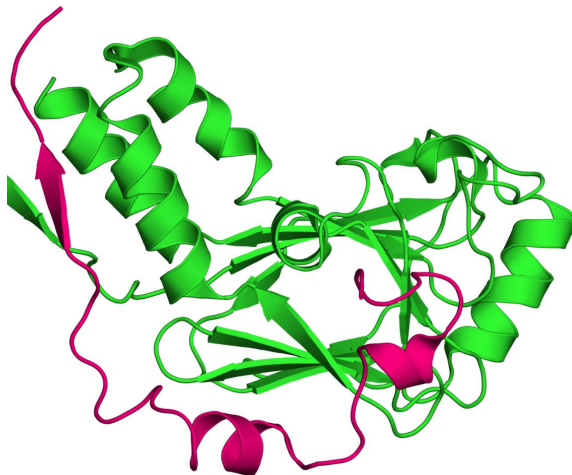
GPCRs

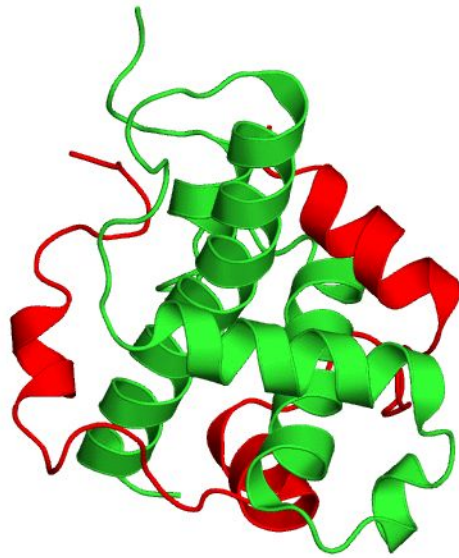
- **Specyficzne właściwości**
- ECD: rozróżnia ligand
- TMD:
 - specyficzna aktywacja w oparciu o motywy *NPxxY* & *CWxP*
 - kontrola allosteryczna (cholesterol)
 - dimeryzacja
- ICD: stabilizacja stanu nieaktywnego poprzez motyw *D/ERY*

Natywnie nie pofałdowane białka (*intrinsically unstructured proteins, IUPs*)

Brak organizacji strukturalnej w białkach

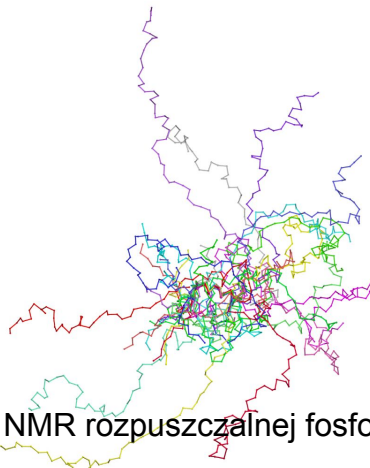
- **Regionalizacja (IDRs)**





Brak organizacji strukturalnej w białkach

- **Całe białka (IUP)**



20 struktur NMR rozpuszczalnej fosfoproteiny



Rola IUPs w
cyklu życiowym komórki

- **Regulacja cyklu komórkowego i ekspresji genów**
- **Pakowanie DNA**
- **Transport komórkowy**
- **Komunikacja i sygnalizacja**
- **Regulacja**
- **Patogeneza (np. nowotwory)**

Metodyka badań IUPs

- **Techniki eksperymentalne**

- ~~Krystalografia~~

- Spektroskopia: NMR/EPR, CD, FRET

- High-speed atomic force microscopy (mikroskopia sił atomowych)

- Raman optical activity

- Time-resolved small angle X-ray scattering (TR-SAXS)

- Metody obliczeniowe:

- IUPred, DISOPRED (prediction)

- Molecular dynamics (characterization)

Wybrane funkcje

- **Regulacja funkcji i procesów biologicznych:**

- Komunikacja

- Sygnalizacja

- Cykle komórkowe

- Ekspresja genów

- **Specyficzne tryby aktywności:**

- rozpoznawanie cząsteczek

- kontrola aktywności i entropii systemu

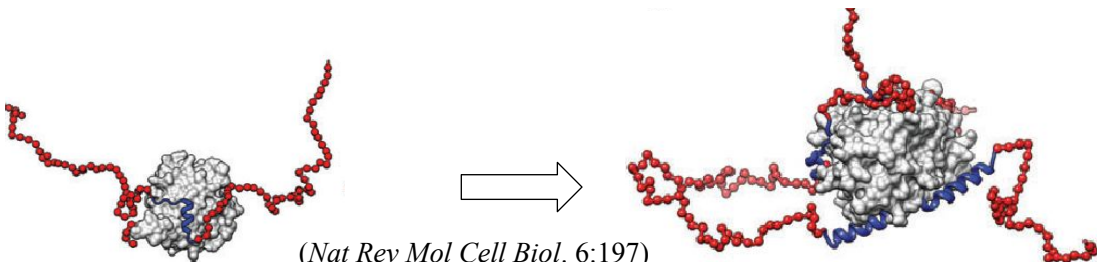
Funkcje

• Rozpoznawanie molekularne

- **Assemblers:** act as scaffolds that spatially direct the assembly of large protein complexes (ribosome, cytoskeleton, chromatin)
- **Effectors:** regulate the activity of other proteins or other regions of the same protein (e.g. p21 and p27 in cell cycle regulation)
- **Scavengers:** store and/or neutralize small ligands (e.g. casein)
- **Display sites (IDRs):** targets for post-translational modifications
- **Chaperones:** assist the folding of proteins and RNA molecules

Rozpoznawanie molekularne

- Wiązanie z partnerem powoduje częściową organizację IUPs
- Pewne fragmenty pozostają elastyczne prawdopodobnie
 - powodując zmiany strukturalne u partnera
 - zmieniając właściwości w miejscu wiązania



Funkcje

- **Kontrola aktywności i entropii systemu**

- IUPs stanowią siłę przeciwną do uporządkowanych zmian strukturalnych lub mogą wpływać na organizację i orientację domen wewnątrz kompleksów białkowych
- Nawet w trakcie aktywności pozostają nie pofałdowane
- Rodzaje:
 - **Sprężyny**: generują pasywną siłę (powrót napiętych mięśni do stanu spoczynkowego)
 - **Separator**: utrzymuje niezbędne odległości
 - **Łącznik**: łączy domeny w nowe konfiguracje

Charakterystyka IUPs

- **Struktura szczątkowa**

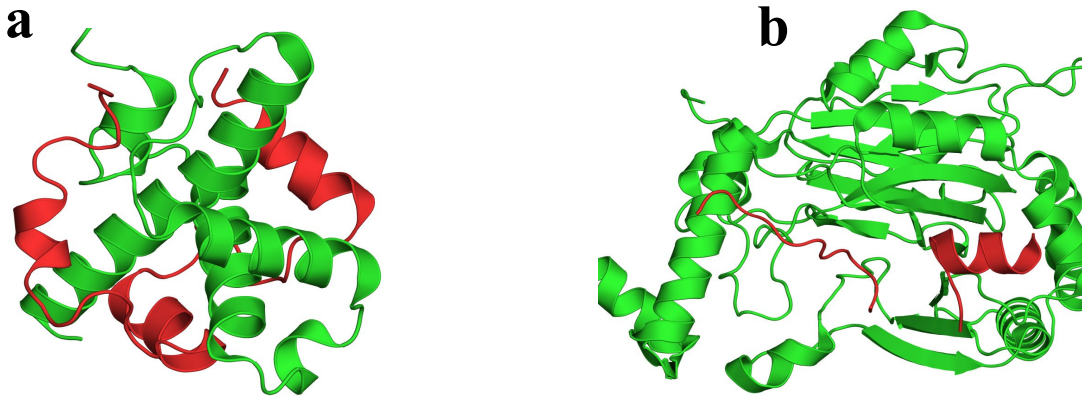
- Helisy **PPII** (polyprolinne II helices) (10-20% aminokwasów)
- Motywy odpowiedzialne za wiązanie:
 - (1) **MoREs/MoRFs** (ang. molecular recognition elements/featurea, 10-70 AA, często amifatyczne),
 - (2) **SLiMs** (ang. short linear motif).

- **Preferencje sekwencji**

- **Pro** → formuje i stabilizuje strukturę drugorzędową
- **Polarne** → blokuje fałdowanie
- **Cys** → brak wiązań disarczkowych
- **SLiMs**: niskie powinowactwo, szybka wymiana
- **Brak konserwacji ewolucyjnej**

Zależność struktura-funkcja IUPs

- **Elastyczność konformacyjna prowadzi do elastyczności wiązania w zależności od warunków.**



The hypoxia-inducible factor-1α (red) bound to different targets

Zależność struktura-funkcja IUPs

- **Powinowactwo wiązania:**

$$\Delta G_{bind} = \Delta H_{bind} - T\Delta S_{bind}$$

□ IUPs → **interakcje wiązanie-faldowanie** → wzrost ΔS_{bind} w porównaniu do białek pofaldowanych → **słabsze wiązanie** →

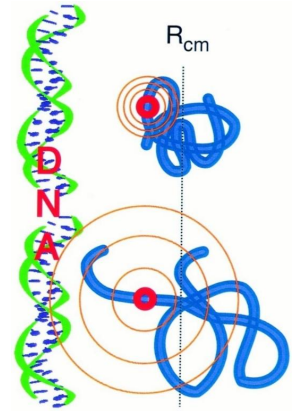
1. **większa częstotliwość wymiany partnerów** (zaleta dla białek sygnalizacyjnych)
2. komponent ΔH_{bind} wyłącza katalizę → **enzymy są ZAWSZE sfaldowane**

□ Niepolarny interfejs → większe ΔH_{bind} kompensuje utratę entropii

Zależność struktura-funkcja IUPs

- **Kinetyka mechanizmu wiązania-fałdowanie**

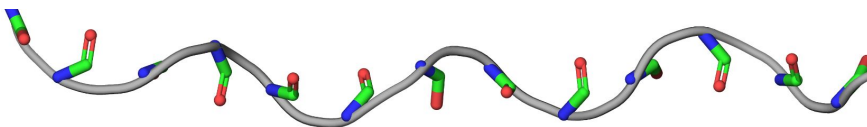
1. Wstępne słabe wiązanie poprzez szcążkowe fałdowanie
2. Wysokie powinowactwo po identyfikacji partnera i osiągnięciu optymalnej konfiguracji.



Zależność struktura-funkcja IUPs

- **Specyficzność substratowa (...lub jej brak)**

- brak specyficznych wiązań wodorowych w strukturze drugorzędowej**
 - **helisy PPII nie posiadają wiązań wodorowych**
- specyficzne wiązanie wykorzystuje dostępne wolne wiązania**



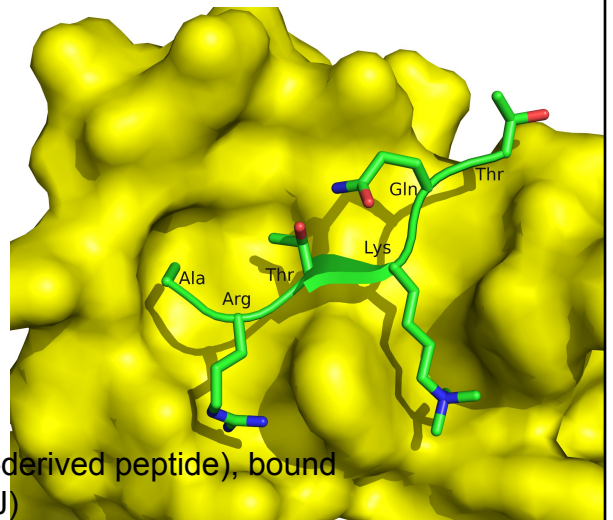
PPII helix with unpaired backbone polar groups

Zależność struktura-funkcja IUPs

- **Reulacja poprzez modyfikacje potranslacyjne**
 - Brak struktury
 - Rozwinięty łańcuch
 - Łatwość enzymatycznej fosforylacji, glikozylacji, acylacji, hydroxylacji...
 - SLiMs są łatwo rozpoznawane przez enzymy PTM

Zależność struktura-funkcja IUPs

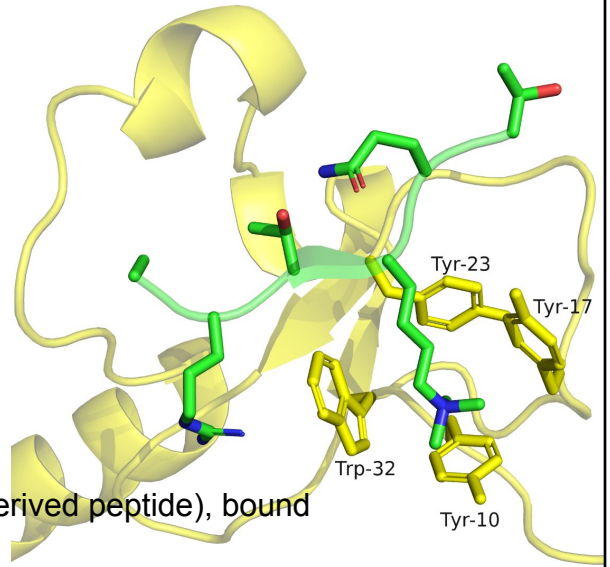
- **Regulacja poprzez modyfikacje potranslacyjne**



Methylated Lys in short linear motif (histone-derived peptide), bound to human BPTF (2F6J)

Zależność struktura-funkcja IUPs

- **Regulacja poprzez modyfikacje potranslacyjne**



Methylated Lys in short linear motif (histone-derived peptide), bound to human BPTF (2F6J)

Metody przewidywania struktury białek

Eksperymentalne metody wyznaczania struktury 3D

- **Metody dyfrakcyjne:**

- X-ray diffraction/scattering
- Neutron scattering
- Electron microscopy/crystallography

- **Metody spektroskopowe**

- Nuclear magnetic resonance (NMR) spectroscopy
- Electron paramagnetic resonance (EPR) spectroscopy
- FTIR, Raman spectroscopy, circular dichroism, mass spectrometry

Przewidywanie struktury

- **Metody eksperymentalne są dokładne ale...**

- są czasochłonne
- są kosztowne
- białka błonowe i kompleksy są kłopotliwe

- Tylko niewielki odsetek białek posiada eksperymentalne struktury.

- **Metody obliczeniowe są mniej dokładne ale szybsze!**

Przegląd stosowanych rozwiązań

- **Podejście fizyczne (*ab initio*)**
 1. Wykorzystanie pól siłowych do obliczenia energii potencjalnej
 2. Próbkowanie konfiguracji
 3. Wykorzystanie teorii pól średnich
- **Podejście oparte o szablony (*comparative*)**
 1. Modelowanie homologiczne
 2. Rozpoznawanie foldu (struktury drugorzędowej)
- **Metody zintegrowane:**
 1. Kierowane eksperymentalnie modelowanie komputerowe
 2. Metody ewolucyjne (skorelowane mutacje)

The Nobel Prize in Chemistry 2013



© Nobel Media AB. Photo: A. Mahmoud
Martin Karplus
Prize share: 1/3



© Nobel Media AB. Photo: A. Mahmoud
Michael Levitt
Prize share: 1/3



© Nobel Media AB. Photo: A. Mahmoud
Arieh Warshel
Prize share: 1/3

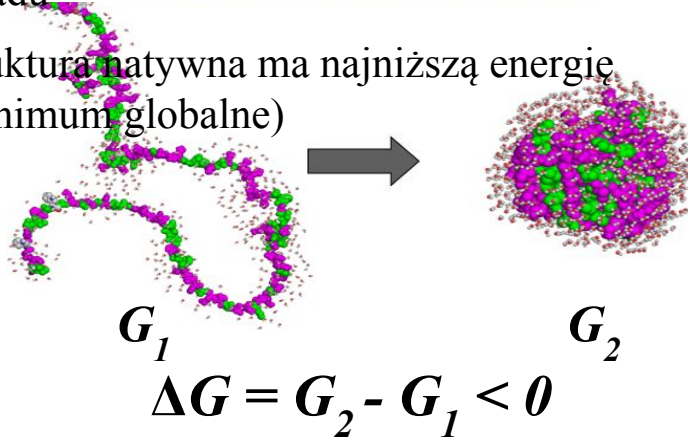
The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel "for the development of multiscale models for complex chemical systems"



Podejście fizyczne

- **Założenia**

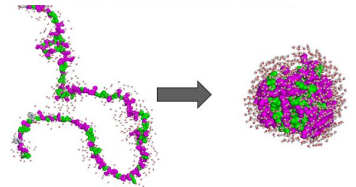
1. Fałdowanie białek bierze się z obniżeniem energii układu
2. Struktura natywna ma najniższą energię (minimum globalne)



Podejście fizyczne

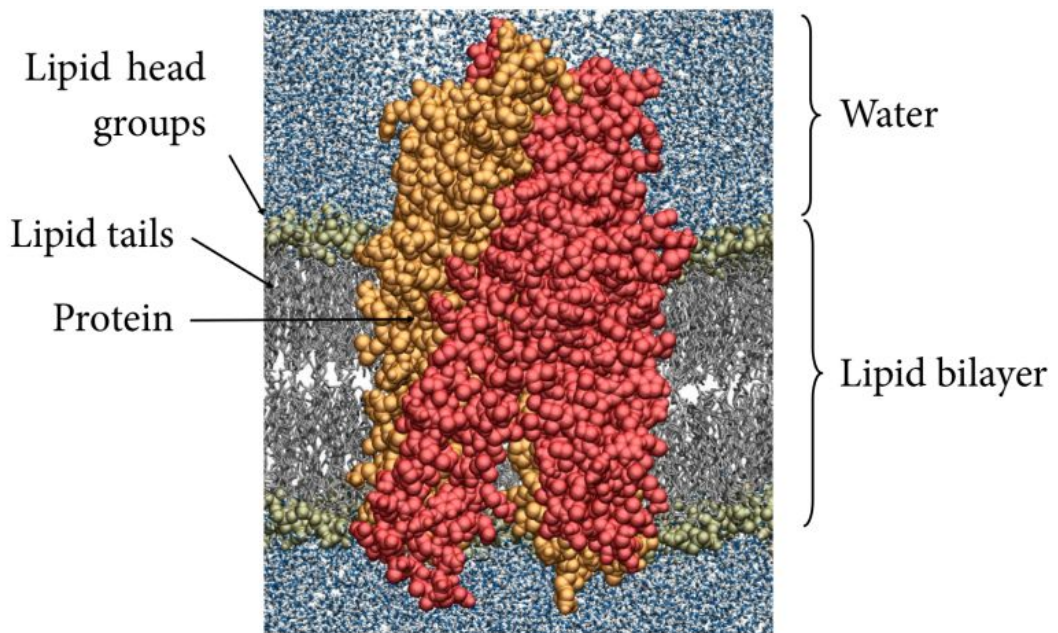
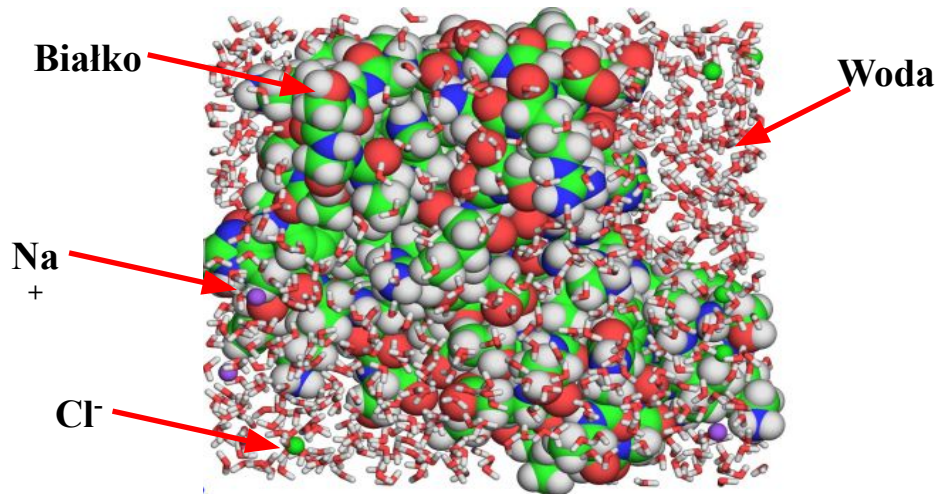
3. Proces fałdowania można opisać poprzez:

- Bezpośrednie (*explicit*) opisanie zmian struktury niepołałdowanej i jej otoczenia (wody, jonów, tłuszczu)
- Matematyczny opis całkowitej energii systemu i konfiguracji (*force field*)
- Algorytm testujący lokalne struktury w celu odzyskania tej o najniższej energii (*natywnej*)



Podejście fizyczne

- Bezpośredni (*explicit*) opis systemu



Podejście fizyczne

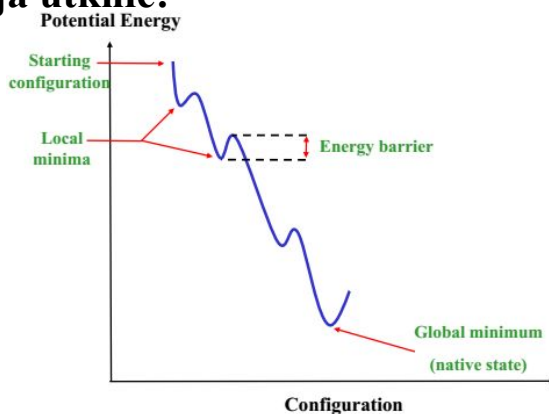
- **Próbkowanie konfiguracji**

- **Opcja 1:** wygenerowanie wszystkich możliwych konfiguracji i identyfikacja optymalnej - oblicznowo niepraktyczne
- **Opcja 2: minimalizacja energii**
 1. Oszacowanie potencjału niepołałdowanego białka
 2. Wprowadzenie lokalnych zmian w lokalizacji atomów i (1)
 3. Niższa energia kontynuuj, brak zmian w energii lub wzrost wróć do (1)

Podejście fizyczne

- **Próbkowanie konfiguracji**

- **Jeżeli symulacja natrafi na lokalne minimum symulacja utknie!**



Metody oparte o szablony

- **Struktura jest generowana w oparciu o szablony (templates), które:**
 1. Wykazują podobieństwo na poziomie sekwencji (homologs)
 2. Posiadają właściwości fizykochemiczne zgodne z analizowaną sekwencją (query)
- **Podstawowe strategie:**
 1. Modelowanie homologiczne
 2. Rozpoznawanie foldu (threading)

Zgodność strukturalne

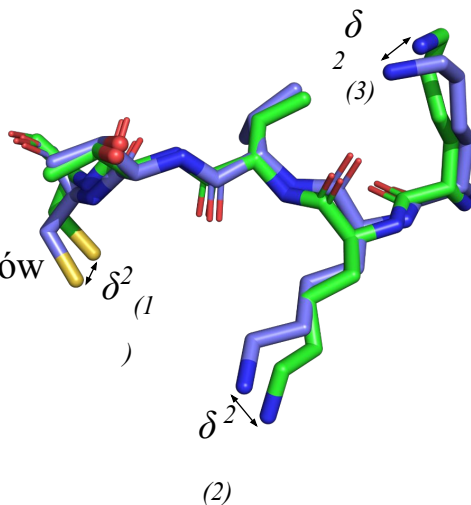
- **R.M.S.D - root-mean-square deviation**

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}$$

□ N - porównywane atomy

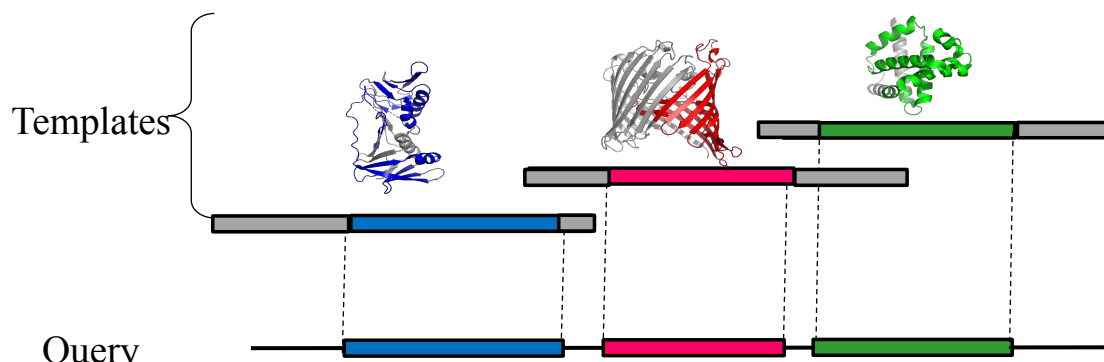
□ δ_i - odległość [Å] pomiędzy parą atomów i :

$$\delta^2 = (x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2$$



Metody oparte o szablony

- Białka o podobnej **sekwencji** mają podobną **strukturę**
- Strukturę białka można przewidzieć po identyfikacji homologii na poziomie sekwencji



Metody oparte o szablony

1. Identyfikacja szablonów > 30% zgodności (psi-BLAST)
2. Dopasowanie sekwencji
3. Integracja współrzędnych zgodnych aminokwasów
4. Optymalizacja energii (minimalizacja)
5. Ocena jakości symulacji (WHATIF, Verify3D)

I. Identyfikacja szablonów

30% zgodności
w dopasowaniu
80 lub więcej
aminokwasów.

```
>THN_DENCL
KSCCPTTAARNQYNICRLPGTFRFVCAALSGCKIISGTGCPFGYRH-
>1crnA
TTCCPSIVARSNFNVCLPGTPEALCATYIGCIIIPGATCPGDYAN-
```


Target	MAVHQIGEGGLVMYVWTFG--LMAFSALAFVMTFRPLNKRSHGYYITLA	48
3am6.1.A	-----ETGMIAQWIVFALM--RRAIAPVAVHFRPELKS-AYYINI	46
Target	IVTIAAIYYAMAASGGKALVSNPDGNLRDIYYARYIDWFFTPPLLLLDI	98
3am6.1.A	AICTIAATAYYMAVINYQDLTNGE---RQVYARYIDW/LTTPLLLLD	93

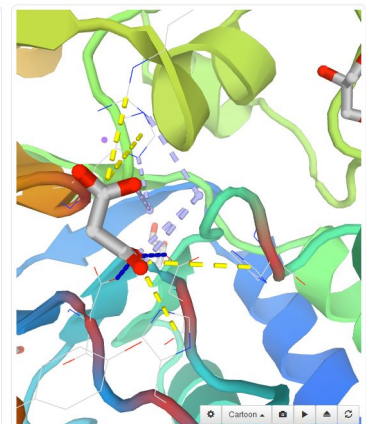
Sequence Identity: 33.80

II. Przygotowanie szablonu

SWISS-MODEL
Template Library
(SMTL)

SMTL ID : 5inb.1
RepoMan-PP1g (protein phosphatase 1, gamma isoform)
holoenzyme complex

Coordinates	PDB Format
Method	X-RAY DIFFRACTION 1.30 Å
Oligo State	hetero-1-1-mer
Ligands	1 x ML1 ¹² : MALONATE ION (Non-functional Binders) ^
	ML1: 6 residues within 4Å ^
	<input checked="" type="checkbox"/> Chain A: D.61, H.63, D.89, R.218, H.245, Y.269
	9 PLIP interactions: ^
	9 interactions with chain A
	<input checked="" type="checkbox"/> Hydrogen bonds: A.D.61, A.D.89
	<input checked="" type="checkbox"/> Water bridges: A.R.218, A.R.218, A.H.245
	<input checked="" type="checkbox"/> Salt bridges: A.H.63, A.H.122, A.R.218, A.H.245
Links	RCSB ¹² PDBe ¹² PDBj ¹² PDBsum ¹² CATH ¹² PLIP ¹²
Citation	Kumar, G.S. et al., The Ki-67 and RepoMan mitotic phosphatases assemble via an identical, yet novel mechanism. <i>Elife</i> (2016) DOI
Release Date	2016-10-05
Peptides	Serine/threonine-protein phosphatase PP1-gamma catalytic subunit: A Cell division cycle-associated protein 2: B
SMTL-PDB	SMTL Chain id: A B



III. Identyfikacja optymalnego szablonu

The screenshot shows the I-TASSER interface with the following details:

- Templates:** Quaternary Structure, Sequence Similarity, Alignment, More
- ITSort:** Coverage, GMQE, QSQE, Identify, Method, Oligo State, Ligands
- Template 1:** 319y.1.A Superoxide dismutase [Cu-Zn]. GMQE: 0.80, QSQE: 0.91, IDDT: 68.83, Method: X-ray, 2.0Å, Oligo State: homo-dimer, Ligands: 2 x Zn²⁺.
- Template 2:** 319e.1.A Superoxide dismutase [Cu-Zn]. GMQE: 0.89, QSQE: 0.92, IDDT: 67.97, Method: X-ray, 2.0Å, Oligo State: homo-dimer, Ligands: 2 x Zn²⁺.
- Template 3 (Selected):** 319y.1.A Superoxide dismutase [Cu-Zn]. GMQE: 0.80, QSQE: 0.92, IDDT: 67.97, Method: X-RAY DIFFRACTION 1.80 Å, Found By: HHblits, Seq Similarity: 0.51, Bionut Oligo State: homo-dimer, Target Prediction: It is possible to build a homo-dimer. The target model is also predicted to be a homo-dimer. Build a homo-dimer (selected) or monomer.
- Build Models:** A 3D ribbon diagram of the protein structure is shown in the 'Build Models' panel.
- Sequence:** Target: `RYVYKAVCVIINGDA...GIVFFDQSSQI...VYSGVCGI...LGGFVRFQDINSG...KSSGDFPFR`

IV. Ocena wyników

GMQE (Global Model Quality Estimate) is a quality estimate which combines properties from the target-template alignment and the template structure. They are combined using a multilayer perceptron trained to predict the IDDT score of the resulting model. The GMQE is available before building an actual model and thus helpful in selecting optimal templates for the modeling problem at hand. Once a model is built, the GMQE (1) in the figure above) gets updated for this specific case by also taking into account the OMEANDISCo global score of the obtained model in order to increase reliability of the quality estimation.

OMEANDISCo global score (Studer et al., (2) in the figure above) is the average per-residue OMEANDISCo score (see below) which has been found to correlate well with the IDDT score (Mariani et al.). The provided error estimate is based on OMEANDISCo global scores estimated for a large set of models and represents the root mean squared difference (i.e. standard deviation) between OMEANDISCo global score and IDDT (the ground truth). As the reliability of the prediction depends on model size, the provided error estimate is calculated based on models of similar size to the input.


OMEAN Z-score analysis (Benkert et al.) is deprecated and the GMQE and OMEANDISCo global scores should be consulted for global model quality estimates instead. It is based on 4 statistical potentials of mean force and their linear combination: the "OMEAN" score. All scores, 5 in total, are compared with what one would expect from experimentally determined structures of similar size using Z-scores (4) in the figure above). In other words: How many standard deviations from the mean is my model score given a score distribution from a large set of experimentally determined structures. Z-scores around 0.0 therefore reflect a "native-like" structure and, as a rule of thumb, a "OMEAN" Z-score below -4.0 indicates a model with low quality. This is illustrated by the "Comparison" plot (5) in the figure above). The x-axis shows protein length (number of residues). The y-axis is the "OMEAN" score. Every dot represents one experimental protein structure. Black dots are experimental structures with a "OMEAN" score within 1 standard deviation of the mean (Z-score) between 0 and 1), experimental structures with a [Z-score] between 1 and 2 are grey. Experimental structure that are even further from the mean are light grey. The actual model is represented as a red star.

The screenshot shows the 'Model Results' page with the following information:

- Order by:** GMQE
- GMQE:** 0.88
- OMEANDISCo Global:** 0.85 ± 0.05
- 1:** GMQE value (0.88)
- 2:** OMEANDISCo Global value (0.85 ± 0.05)
- 3:** OMEANDISCo Local plot (Local Quality Estimate)
- 4:** OMEAN Z-Scores plot (Comparison plot)
- 5:** OMEAN Z-Scores plot (Comparison plot)
- Model 01:** Structure Assessment
- Template:** 319y.1.A, Seq Identity: 67.97%, Coverage: [bar chart]
- Description:** Superoxide dismutase [Cu-Zn]. Crystal structures of holo and Cu-deficient CuZnSOD from the silkworm Bombyx mori and the implications in Amyotrophic lateral sclerosis
- Model-Template Alignment:**

Model_01:1A	RYVYKAVCVIINGDA...GIVFFDQSSQI...VYSGVCGI...LGGFVRFQDINSG...KSSGDFPFR	80
Model_01:1B	RYVYKAVCVIINGDA...GIVFFDQSSQI...VYSGVCGI...LGGFVRFQDINSG...KSSGDFPFR	80
319y.1.A	RYVYKAVCVIINGDA...GIVFFDQSSQI...VYSGVCGI...LGGFVRFQDINSG...KSSGDFPFR	80

Single-sequence protein structure prediction using a language model and deep learning

Ratul Chowdhury^{1,8}, Nazim Bouatta^{1,8} , Surojit Biswas^{2,3,8}, Christina Floristean^{4,8}, Anant Kharkar⁴, Koushik Roy⁴, Charlotte Rochereau⁵, Gustaf Ahdrizt⁶, Joanna Zhang⁴, George M. Church^{1,2}, Peter K. Sorger^{1,7}  and Mohammed AlQuraishi^{4,6} 

AlphaFold2 and related computational systems predict protein structure using deep learning and co-evolutionary relationships encoded in multiple sequence alignments (MSAs). Despite high prediction accuracy achieved by these systems, challenges remain in (1) prediction of orphan and rapidly evolving proteins for which an MSA cannot be generated; (2) rapid exploration of designed structures; and (3) understanding the rules governing spontaneous polypeptide folding in solution. Here we report development of an end-to-end differentiable recurrent geometric network (RGN) that uses a protein language model (AminoBERT) to learn latent structural information from unaligned proteins. A linked geometric module compactly represents C_α backbone geometry in a translationally and rotationally invariant way. On average, RGN2 outperforms AlphaFold2 and RoseTTAFold on orphan proteins and classes of designed proteins while achieving up to a 10⁴-fold reduction in compute time. These findings demonstrate the practical and theoretical strengths of protein language models relative to MSAs in structure prediction.

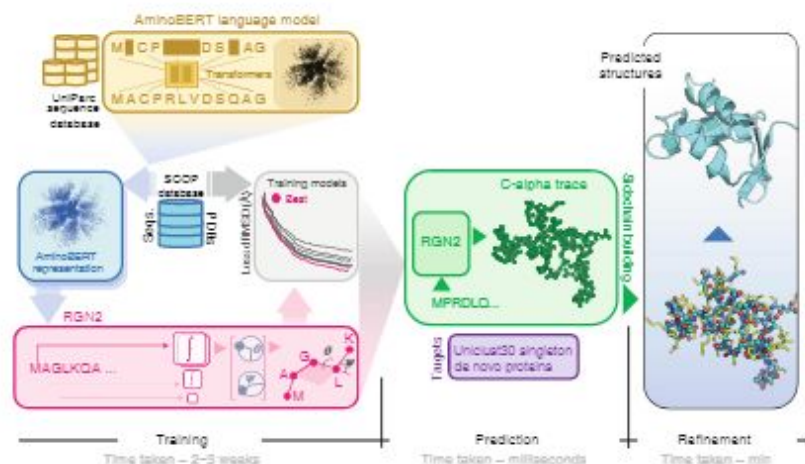


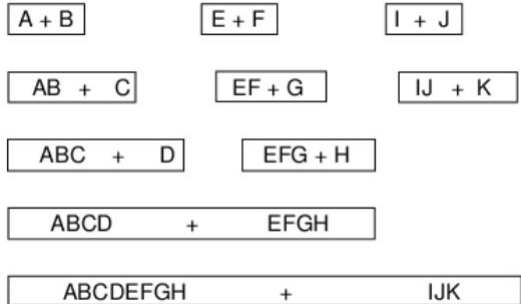
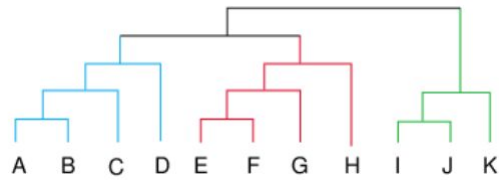
Fig. 1 | Organization and application of RGN2. RGN2 combines a transformer-based protein language model (AminoBERT, yellow) with an RGN that uses Frenet-Serret frames to generate the backbone structure of a protein (green). After initial construction of the sidechains and hydrogen-bonded networks, refinement of the structure is subsequently performed using AF2Rank (blue).

drzewo przewodnie

Dla zbioru wszystkich sekwencji poddawanych dopasowaniu generowane są przyrównania lokalne.

Algorytm działa na zasadzie obliczania wyników podobieństwa jako liczby dopasowań k-tup między dwiema sekwencjami, biorąc pod uwagę ustaloną karę za przerwy. Im bardziej podobne sekwencje, tym wyższy wynik, im bardziej rozbieżne, tym niższe wyniki. Gdy sekwencje zostaną ocenione, generowany jest dendrogram w celu przedstawienia kolejności dopasowania wielu sekwencji.

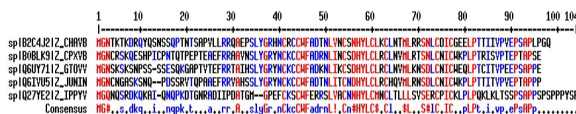
Wyższe uporządkowane zestawy sekwencji są wyrównane jako pierwsze, a następnie pozostałe w kolejności malejącej.



msa można wykonać na więcej niż jeden sposób

Do dopasowania wielu sekwencji można wykorzystać wiele programów:

- CLUSTAL Omega
- MAFFT
- MUSCLE
- T-Coffe

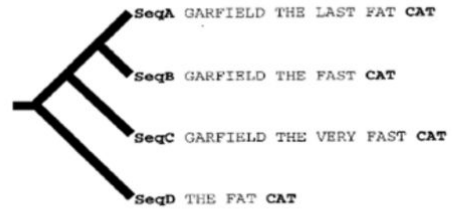


Uzyskane wyniki co do zasady będą porównywalne ale każdy z programów ze względu na specyficzny algorytm może wprowadzać różnice w ostatecznym przyrównaniu. Również prezentacja wyników będzie dla każdego programu inna.

czy kolejność wprowadzenia sekwencji w dopasowaniu będzie mieć znaczenia?

Im więcej sekwencji tym większy ,ciężar' poprawnie dopasowanych fragmentów sekwencji.

Bez interwencji badacza automatyczna korekta dopasowanie jest niezwykle trudna.



```

SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD ----- THE ---- FA-T CAT
    
```

jak postępować...

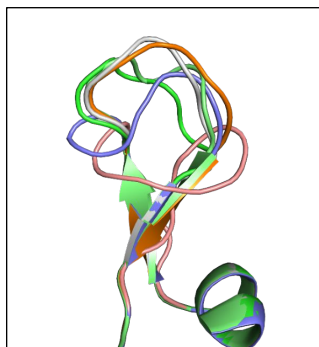
- Im więcej sekwencji tym lepiej...
- Im mniej podobne sekwencje tym lepiej...
- Im bardziej zgodna pod względem długości pula sekwencji tym bardziej wiarygodne dopasowanie...

Input data	Recommendation
A small number of unusually long sequences (>20,000 residues)	Use CLUSTALW. Other program may run out of memory, causing an abort.
2-100 sequences of typical protein length (maximum around 10,000 residues)	Use T-Coffee, MAFFT or MUSCLE
100-500 sequences that are globally alignable	Use MUSCLE or MAFFT
>500 sequences	Use MUSCLE or MAFFT with a faster option

Modelowanie homologiczne

- **Problemy**

1. Liczba dostępnych szablonów jest **ograniczona**.
2. Elementy dynamiczne (pętle, zwroty) nie są dobrze zakonserwowane pomiędzy homologami.



Modelowanie homologiczne

- **Rozpoznawanie foldów**

- Białka o podobnej strukturze posiadają wspólne cechy na poziomie sekwencji
- Cech można zidentyfikować w oparciu o analizę statystyczną
-
- Białko o znanej strukturze może służyć za wzorzec gdy posiada zgodność na poziomie sekwencji

Modelowanie homologiczne

- **Identyfikacja cech**

1. **Na poziomie aminokwasowym:**

- kąty
- powierzchnia
- struktura drugorzędowa
- etc.

2. **Analiza statystyczna**

- dopasowanie wielu sekwencji ujawnia statystyczne trendy

Modelowanie homologiczne

- **Identyfikacje foldu:**

1. Pozycja aminokwasu w sekwencji zapytania kodowana jest tak by odzwierciedlać tendencję (profil) zmian w aminokwasach.
2. Profil analizowany jest w odniesieniu do biblioteki struktur.
3. Pozytywne dopasowanie oznacza identyfikację foldu.
4. Brak dopasowania = nowy fold.

- **Software/servers:**

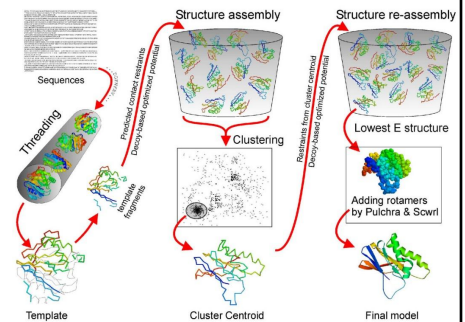
GenTHREADER, SPARKS-X, HHPred, TASSER

Modelowanie homologiczne

- **Metody zintegrowane**

- Każde podejście do analizy jest dobre!
- Np.: HM identyfikacja modeli, minimalizacja energii do optymalizacji.
- Software/servers: **I-TASSER**, **Rosetta**

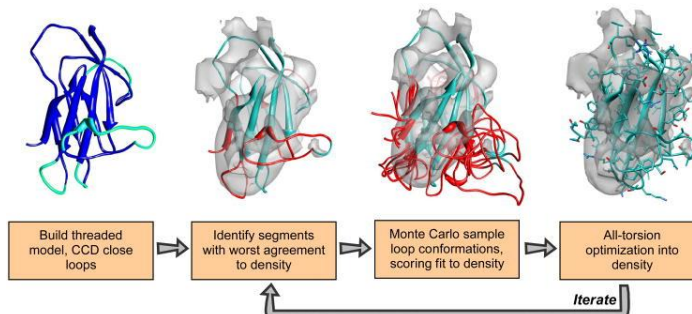
Nature Protocols
(2010) 5 725–738



Modelowanie homologiczne

- **MOdelowanie wspierane eksperymentem**

- Dane strukturalne o niskiej rozdzielczości (**EM, SAXS**) stanowią wartościowe ograniczenia w porociesie modelowania



J Mol Biol. (2009) 392:
181-90

A Rosetta implementation with cryo-EM constraints

Bazy danych modeli białek

- The protein model portal (PMP):

<http://www.proteinmodelportal.org/>

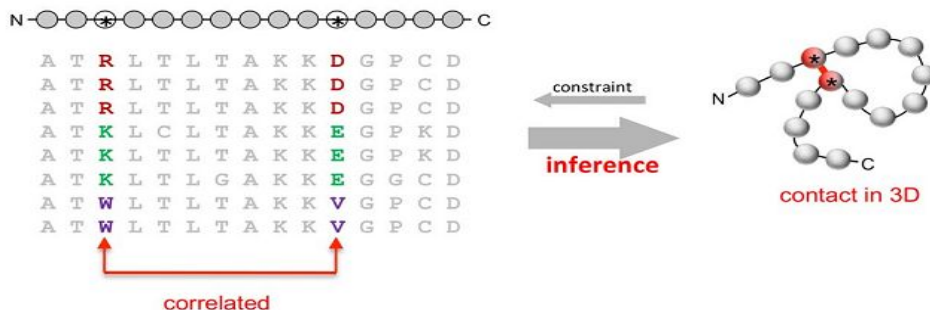
- The protein model database (PMDB):

<https://bioinformatics.cineca.it/PMDB/>

Evolutionary methods

- Prediction by correlated mutations

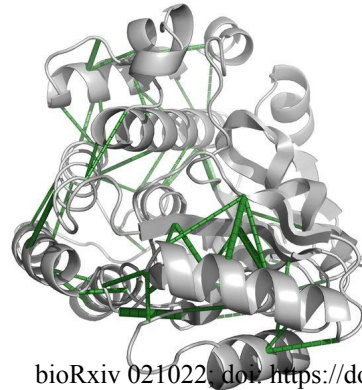
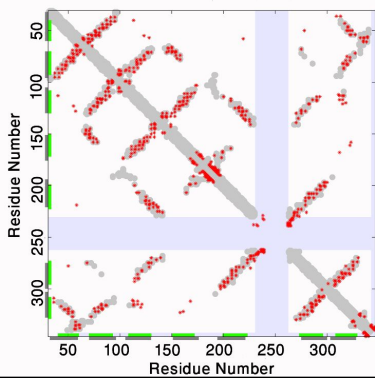
- Positions that are close in space tend to co-evolve
- Therefore, data on **co-evolving positions** can be used as **distance constraints** on structure prediction



Evolutionary methods

- **Prediction by correlated mutations**

- The sequence information is used to build a **contact map**, which is integrated into a **structure prediction algorithm**.



bioRxiv 021022 | doi: <https://doi.org/10.1101/021022>

Evolutionary methods

- **Prediction by correlated mutations**

- **Web servers:** EVcouplings, RaptorX-Contacts, GREMLIN, DESTINI

The screenshot shows the EVcouplings web server interface. It features a large text input field for the protein sequence, with a placeholder text: "UniProt accession, UniProt entry name, sequence in FASTA format, or raw protein sequence." Below the input field is a blue button labeled "Please enter your protein sequence". To the right of the input field is a link for "Examples". Below the input field is a link for "Show advanced settings". At the bottom, there are optional input fields for "Job name (optional)" and "Your e-mail address (optional)".

The screenshot shows the RaptorX Contact Predict web server interface. It features a header with the RaptorX logo and the text "Contact Predict". Below the header is a navigation bar with links for "New Job", "Job Status", "My Jobs", and "Inquiry & Support". The main heading is "Distance-based Protein Folding Powered by Deep Learning". Below the heading is a paragraph of text describing the server's capabilities: "This web server predicts local structure properties, contact and distance matrix, inter-residue orientation and tertiary structure of a protein using an ultra-deep convolutional residual neural network from primary sequence or a multiple sequence alignment. It works particularly well on proteins without many sequence homologs. This server was ranked 1st in contact prediction in both CASP12 and CASP13. See [here](#) for an official ranking list published by the CASP12 assessors and 1, 2 for the method description. Currently the server will predict contact/distance matrix for a protein of no more than 1300 amino acids, and 3D models for a protein of no more than 1000 amino acids." Below the text is a "Job Identification" section with input fields for "Jobname" and "Email". Below that is a "Sequence for Prediction" section with a link for "Example" and a checkbox for "check here if the input is a single Multiple Sequence Alignment (jobs and format)".