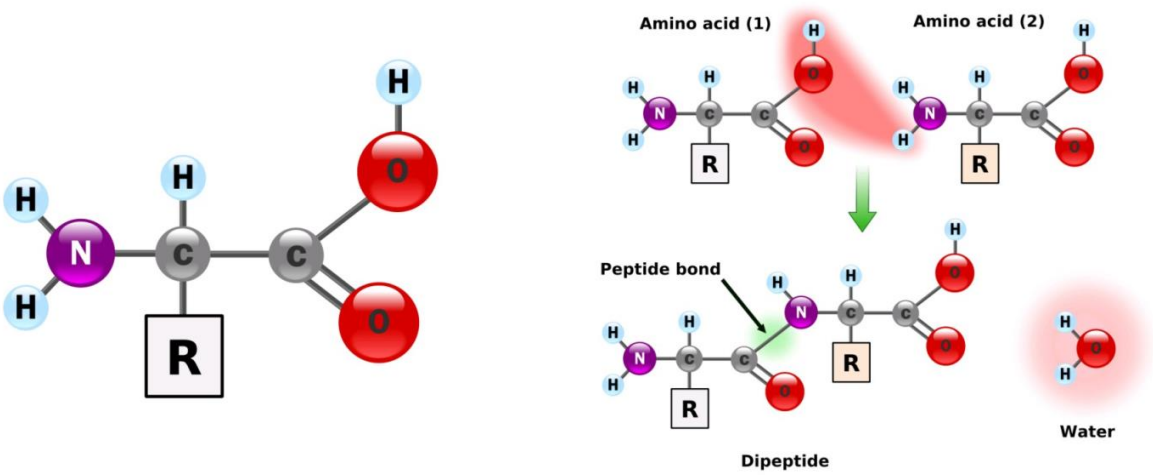
Instytut Informatyki i Matematyki Komputerowej UJ,
opracowanie: mgr Ewa Matczyńska, dr Jacek Śmietański

Aminokwasy

1. Budowa aminokwasów

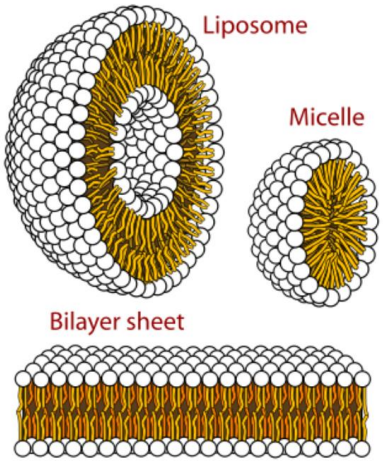


Rysunek 1: Budowa aminokwasów i białek.

- Aminokwasy są związkami zawierającymi zasadową grupę aminową -NH2 oraz kwasową grupę karboksylową –COOH.
- Obie te grupy przyłączone są do tzw. węgla *Cα*, do którego przyłączony jest również atom wodoru oraz grupa R, czyli łańcuch boczny.
- Aminokwasy różnią się od siebie budową łańcucha bocznego, jest on elementem charakterystycznym dla danego typu aminokwasu.
- Aminokwasy łączą się tzw. wiązaniami peptydowymi – grupa aminowa jednego aminokwasu łączy się z grupą karboksylową drugiego aminokwasu, w wyniku wiązania powstaje cząsteczka wody.
- Ze względu na właściwości łańcucha bocznego amiokwasy można podzielić na grupy (rys. poniżej).
- Wyróżniamy aminokwasy które posiadają ładunek elektryczny („+” np. R,H,K; „-” np. D,E).
- Te aminokwasy, które nie mają ładunku również mogą wykazywać właściwości, tak jakby częściowo posiadały ładunek. Na skutek obecności bardzo elektroujemnego

pierwiastka np. azotu, chmura elektronów może przemieścić się w stronę tego pierwiastka tworząc wiązanie spolaryzowane, dlatego aminokwasy te nazywamy polarnymi (S, T, N, Q).

- Cząsteczki wody również mają spolaryzowane wiązania, gdyż elektrony przesuwają się w stronę tlenu, gdzie zgromadzony jest częściowy ładunek ujemny. Ponieważ woda jest również substancją polarną, aminokwasy posiadające ładunek oraz aminokwasy polarne, dobrze łączą się z wodą – są hydrofilowe (z gr. *hydros* - woda i *philia* - miłość, lubiące wodę).



Rysunek 2: Struktury lipidowe starają się zminimalizować kontakt części hydrofobowych z wodą, przybierają postać dwuwarstwowej błony lub miceli.

- Przeciwieństwem aminokwasów hydrofilowych są aminokwasy hydrofobowe (z gr. *phobos* – strach, których łańcuchy boczne składają się z łańcuchów węglowodorowych, bądź pierścieni aromatycznych (A, I, L, M, F, W, Y, V). Kontakt ich powierzchni z wodą jest energetycznie (właściwie entropijnie) niekorzystny. Można wyobrażać sobie hydrofobowe aminokwasy jako cząsteczki tłuszczu w wodzie, które łączą się, aby zminimalizować powierzchnię kontaktu z wodą.
- Aminokwasy wyróżnione na rysunku poniżej jako specjalne, mają rzeczywiście ciekawe właściwości:

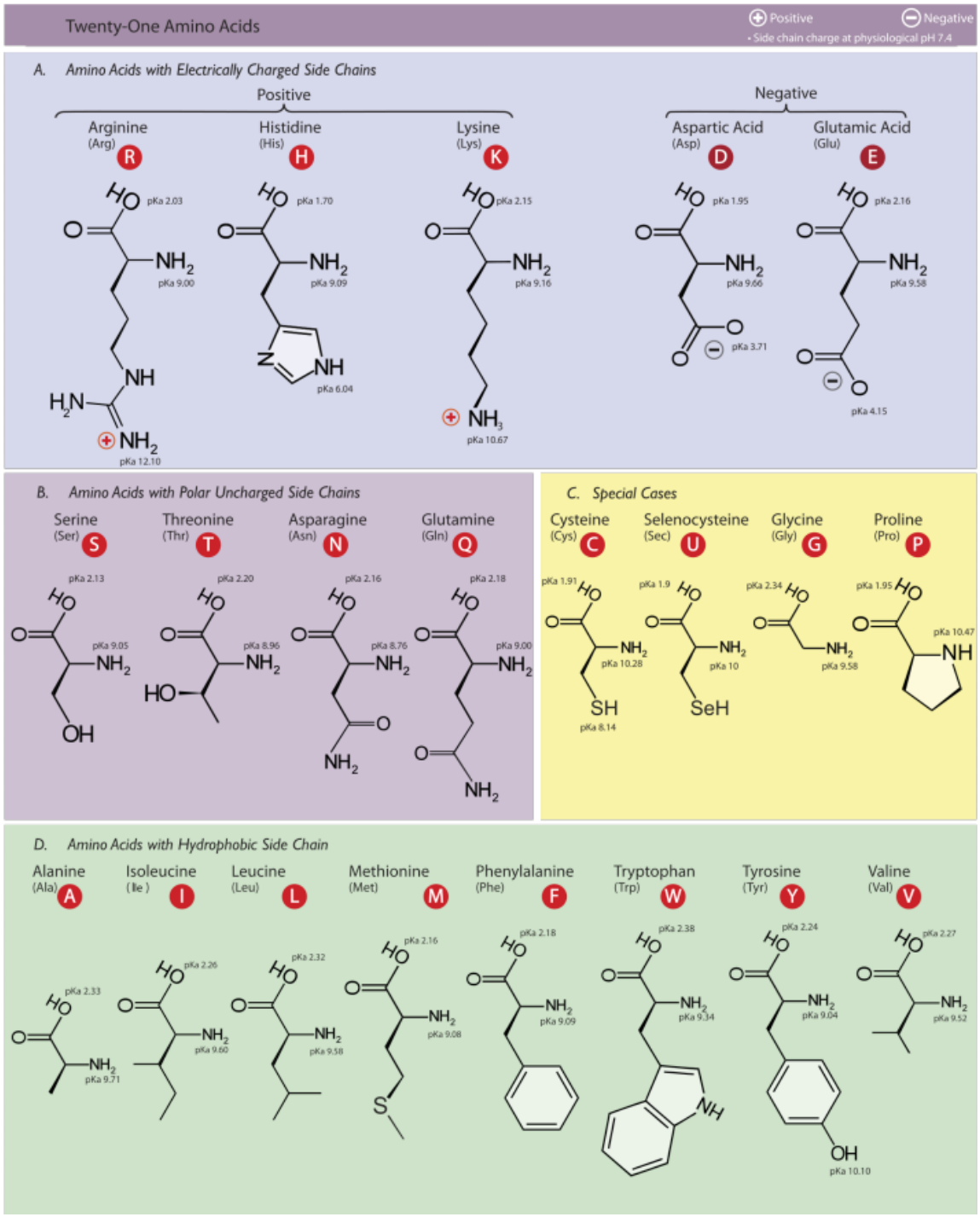
  - glicyna (G) jako łańcuch boczny ma tylko wodór, jest najmniejszym aminokwasem,

  - prolina (P) ma badzo sztywną strukturę ze względu na pierścień aromatyczny, w którym zawiera się grupa aminowa,

- cysteina (C), ze względu na atom siarki może tworzyć tzw. wiązania dwusiarczkowe w strukturze białka,

W zasadzie G i P możemy zaliczyć do aminokwasów hydrofobowych, natomiast C do aminokwasów polarnych.

Będziemy starali się wyodrębnić te grupy za pomocą analizy składowych głównych.



Rysunek 3: Grupy i struktury poszczególnych aminokwasów.

2. PCA – analiza składowych głównych

Analiza składowych głównych polega na poszukiwaniu wyjaśnienia korelacyjej struktury zbioru zmiennych przy użyciu mniejszego zbioru kombinacji liniowych tych zmiennych. Te liniowe kombinacje są nazywane składowymi głównymi. Inaczej mówiąc całkowitą zmienność zbioru danych składającego się z m zmiennych można często zachować dla mniejszego zbioru k nowych zmiennych, będących liniowymi kombinacjami zmienych pierwotnych. Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować wariancję wzdłuż kolejnych współrzędnych. W ten sposób konstruowana jest nowa przestrzeń obserwacji, w której najwięcej zmienności wyjaśniają początkowe składowe.

Załóżmy że mamy macierz $X \in R^{n \times m}$, n obserwacji z przestrzeni m wymiarowej. Jak wyznaczyć składowe główne?

1. Na początku można ustandaryzować zmienne, czyli doprowadzić do sytuacji kiedy każda z m zmiennych będzie mała średnią $\mu = 0$ oraz odchylenie stamdardowe $\sigma = 1$

   $$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

   po to aby każda zmienna mogła mieć propocjonalny wpływ na obliczenie składowych.
2. Obliczamy macierz kowariancji, która opisuje kowariancję, czyli zależność liniową między zmiennymi:

   $$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{bmatrix}$$

   gdzie: $\sigma_{ij}^2 = \frac{\Sigma_{k=1}^{n}(x_{ki}-\mu_i)(x_{kj}-\mu_j)}{n}$

3. Okazuje się, że wektor współczynników *i-tej* głównej składowej v_i jest równy wektorowi własnemu macierzy kowariancji Σ odpowiadającemu wartości własnej λ_i, gdzie $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ są wartościami własnymi Σ.
   Szukamy, więc rozkładu macierzy $\Sigma = V\ D\ V^T$, gdzie V jest macierzą wektorów własnych (wektory własne w kolumnach), a D macierzą z wartościami własnymi na

przekątnej. Ponieważ Σ jest macierzą symetryczną to wiemy, że jej wartości własne są rzeczywiste, a wektory własne są ortogonalne.

4. Następnie wystaczy przedstawić wektory obserwacji w nowej bazie złożonej z wektorów własnych macierzy Σ:

$$x'_j = V^T x_j; j = 1 \ldots n$$

Możemy ograniczyć liczbę wektorów i rzutować do przestrzeni niżej wymiarowej $k<m$. Sumując wartości własne odpowiadające k wektorom własnym, które wybraliśmy możemy się dowiedzieć jaką częśc wariancji oryginalnego zbioru danych zachowujemy po takim rzutowaniu.

Zadanie:

Badanie właściwosci aminokwasów (4 pkt)

Rozwiązanie zadanie prześlij mailem do wtorku, **21.01.2020** włącznie, na adres:
jacek.smietanski@ii.uj.edu.pl
Temat wiadomości proszę opatrzyć przedrostkiem **[Bio] Lab 12**. Rozwiązaniem ma być **tylko jeden plik** – skrypt zgodny z Pythonem w wersji 3.x**,** zawierający wszystkie niezbędne funkcje oraz procedurę wykonawczą. Proszę o nazwanie pliku wg schematu**: Imie.Nazwisko.12.py**.

Stosując analizę głównych składowych, postaramy się zobrazować właściwości aminokwasów. Tabela poniżej przedstawia różne właściwości poszczególnych aminokwasów.

	V	B	P	pI	H1	H2	SAS	FA
A	67	11,5	0	6	1,8	1,6	113	0,74
R	148	14,28	52	10,76	-4,5	-12,3	241	0,64
N	96	12,28	3,38	5,41	-3,5	-4,8	158	0,63
D	91	11,68	49,7	2,77	-3,5	-9,2	151	0,62
C	86	13,46	1,48	5,05	2,5	2	140	0,91
Q	114	14,45	3,53	5,65	-3,5	-4,1	189	0,62
E	109	13,57	49,9	3,22	-3,5	-8,2	183	0,62
G	48	3,4	0	5,97	-0,4	1	85	0,72
H	118	13,69	51,6	7,59	-3,2	-3	194	0,78
I	124	21,4	0,13	6,02	4,5	3,1	182	0,88
L	124	21,4	0,13	5,98	3,8	2,8	180	0,85
K	135	15,71	49,5	9,74	-3,9	-8,8	211	0,52
M	124	16,25	1,43	5,74	1,9	3,4	204	0,85
F	135	19,8	0,35	5,48	2,8	3,7	218	0,88
P	90	17,43	1,58	6,3	-1,6	-0,2	143	0,64
S	73	9,47	1,67	5,68	-0,8	0,6	122	0,66
T	93	15,77	1,66	5,66	-0,7	1,2	146	0,7
W	163	21,67	2,1	5,89	-0,9	1,9	259	0,85
Y	141	18,03	1,61	5,66	-1,3	-0,7	229	0,76
V	105	21,57	0,13	5,96	4,2	2,6	160	0,86
mean	109,2	15,35	13,59	6,03	-0,5	-1,4	175	0,74

Rysunek 4: Właściwości aminokwasów.

1. ***V*** - objętość aminokwasu.
2. ***B*** - tęgość aminokwasu, czyli stosunek objętości łańcucha bocznego do jego długości.
3. ***P*** - indeks polarności (bierze pod uwagę ładunek elektryczny jak i polaryzację wiązań, ale nie rozróżnia znaku ładunku).
4. ***pI*** - pH punktu izoelektrycznego: rozróżnia ładunek aminokwasów. Aminokwasy z ujemnym ładunkiem będą miały niższe *pI*, czyli są bardziej kwasowe. Aminokwasy z dodatnim ładunkiem są bardziej zasadowe.
5. ***H1, H2*** - dwie skale hydrofobowości, im wyższa wartość, tym bardziej hydrofobowy aminokwas.
6. ***SAS*** - pole powierzchni dostępnej dla wody w rozwiniętym łańcuchu aminokwasów.
7. ***FA*** - ułamek pola powierzchni, który w procesie fałdowania białka staje się dla wody niedostępny.

Dla danej tabeli właściwości aminokwasów (plik *aaProperties.txt*) użyj analizy głównych składowych do zaobserwowania właściwości aminokwasów:

- oblicz główne składowe (w pakiecie *numpy* mamy odpowiednie funkcje : *mean, std, eig*);

- narysuj wykres w dwóch wymiarach względem pierwszej i drugiej głównej składowej (używając np. pakietu *matplotlib*): wyświetl ile wariancji wyjaśniają dwie pierwsze składowe;
- zinterpretuj pozycje poszczególnych aminokwasów na wykresie, biorąc pod uwagę udział poszczególnych, oryginalnych zmiennych w głównych składowych, oraz położenie aminokwasów względem siebie.