BIOINFORMATYKA

edycja 2018 / 2019

wykład 6

Dopasowanie par sekwencji
Metody heurystyczne

dr Jacek Śmietański
jacek.smietanski@ii.uj.edu.pl
http://jaceksmietanski.net

Plan wykładu

1. Statystyczna istotność dopasowań
2. Metody heurystyczne – wprowadzenie
3. Algorytm BLAST
4. BLAST w praktyce

Statystyczna istotność dopasowań

Dlaczego o tym mówimy?

Czy znalezione dopasowanie sekwencji jest przypadkowe, czy może rzeczywiście świadczy o ewolucyjnym pokrewieństwie sekwencji?

inaczej:

Czy potrafimy wskazać jaka wartość optymalnego dopasowania świadczyłaby o homologii, a jaka jedynie o przypadkowym podobieństwie?

Chcemy zatem zdefiniować miarę, która będzie nam wskazywać **istotność dopasowania.**

Ocena istotności dopasowania

1. Tworzenie zestawu losowych sekwencji, o tej samej długości i składzie co rzeczywiste.

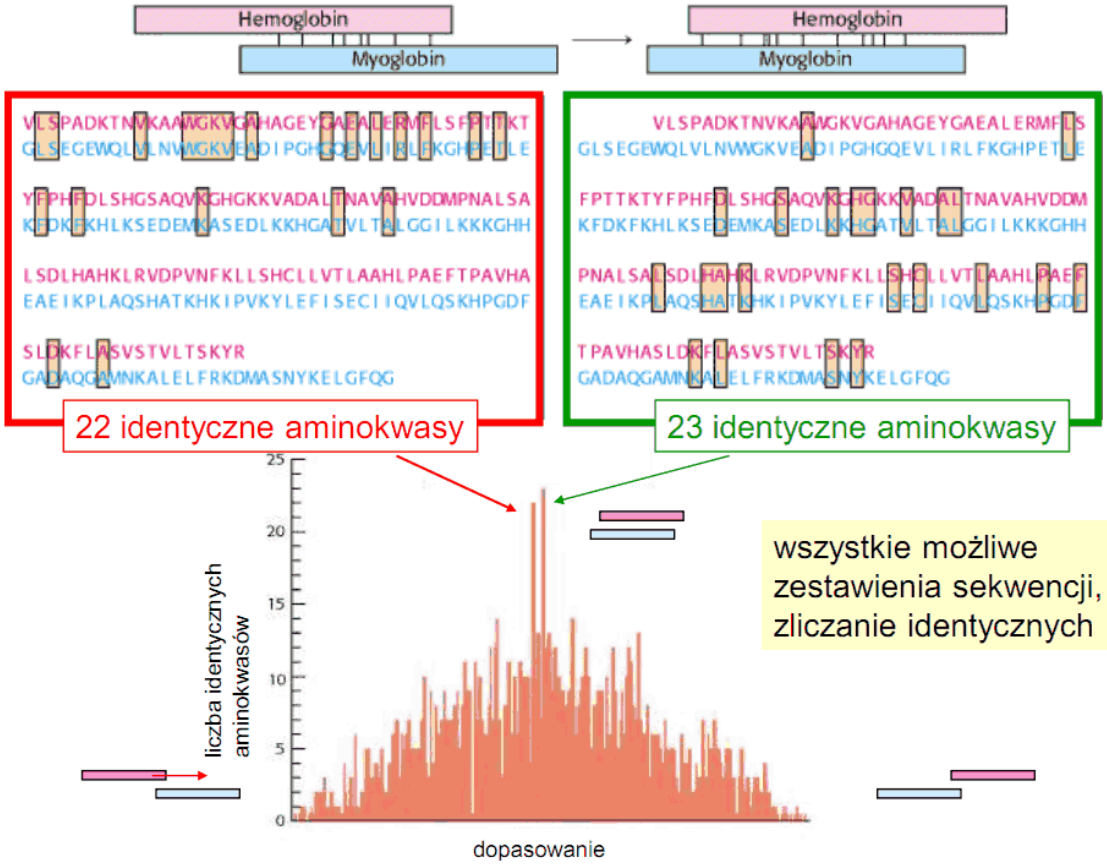
2. Przyrównanie parami wygenerowanych losowych sekwencji przy tych samych parametrach dopasowania.

3. Określenie rozkładu punktacji, średniej i odchylenia standardowego (SD).

Rozkład punktacji nie jest rozkładem normalnym.

Jakość dopasowania globalnego

Przykład:
Porównanie obliczonej wartości dla danego dopasowania z wartościami obliczonymi dla wielu dopasowań przypadkowych sekwencji o podobnym składzie i długości.



dopasowanie sekwencji α hemoglobiny ludzkiej i mioglobiny ludzkiej

Jakość dopasowania, obliczanie

Drugi przykład:
Zestawienia sekwencji z uwzględnieniem przerw, zliczanie identyczności w dopasowaniach.

Optymalne dopasowanie:
38 identycznych aminokwasów we fragmencie o długości 148,
tj. 25.9% identycznych aminokwasów.

Czy jest to znaczące podobieństwo?

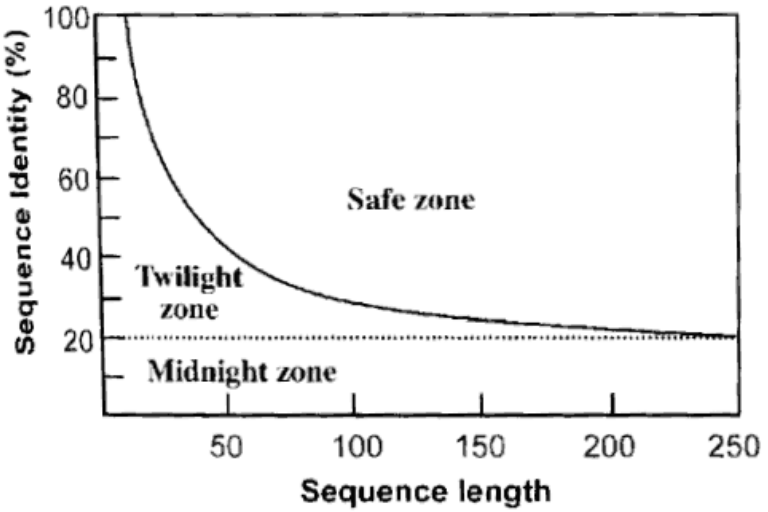
Jakość dopasowania - praktyka

Przykładowo, dla sekwencji o długości 100 aminokwasów:

• identyczność wyższa niż 30% prawie na pewno nie jest przypadkowa – przyjmujemy, że jest homologia;

• identyczność niższa niż 20%: mało prawdopodobne, żeby podobieństwo było biologicznie istotne

Łatwo ocenić istotność statystyczną.
Trudno ocenić istotność biologiczną.
⇒ Brak statystycznej istotności nie wyklucza homologii.



ch.EMBnet.org

Home | Services | Courses | Links | Contacts

PRSS3 - evaluates the significance of a protein sequence alignment

prss3 is used to evaluate the significance of a protein or DNA sequence similarity score by comparing two sequences and calculating optimal similarity scores, and then repeatedly shuffling the second sequence, and calculating optimal similarity scores using the Smith-Waterman algorithm. An extreme value distribution is then fit to the shuffled-sequence scores. The characteristic parameters of the extreme value distribution are then used to estimate the probability that each of the unshuffled sequence scores would be obtained by chance in one sequence, or in a number of sequences equal to the number of shuffles.
This program is derived from rdf2, described by Pearson and Lipman, PNAS (1988)

Obliczane istotności zestawienia sekwencji
http://www.ch.embnet.org/software/PRSS_form.html

within a local window, so that the order of residues 1-10, 11-20, etc, is destroyed but a residue in the first 10 is never swapped with a residue outside the first ten, and so on for each local window.
This program is part of the FASTA package of sequence analysis program. The complete package is available by anonymous ftp from ftp.virginia.edu.

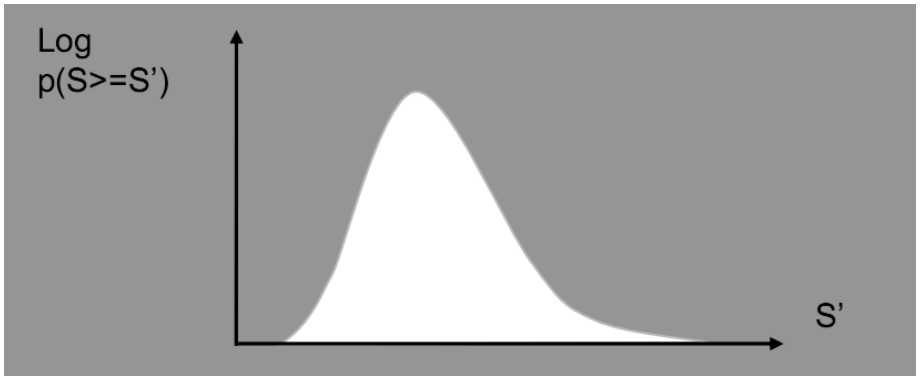
Usage: Paste your two sequences in one of the supported formats into the sequence fields below
and press the "Run PRSS" button.
Make sure that both format buttons (next to the sequence fields) shows the correct formats

Number of shuffles :	200 window size: 10
Scoring matrix :	default
gap opening penalty:	12 gap extension penalty: 2
First sequence title (optional):	
Input	

Istotność statystyczna dopasowania lokalnego

Dla dopasowań lokalnych rozkład maksymalnych wartości punktacji dopasowania dla sekwencji losowych przyjmuje rozkład wartości ekstremalnych, *extreme values distribution* (Karlin i Altschul 1990).

S – punktacja podobieństwa
m – długość porównywanej sekwencji
n – wielkość bazy
λ – parametr określajacy wpływ systemu punktowania
K – liczba powtarzających się segmentów (nukleotydów/aminokwasów) w przeszukiwanej sekwencji



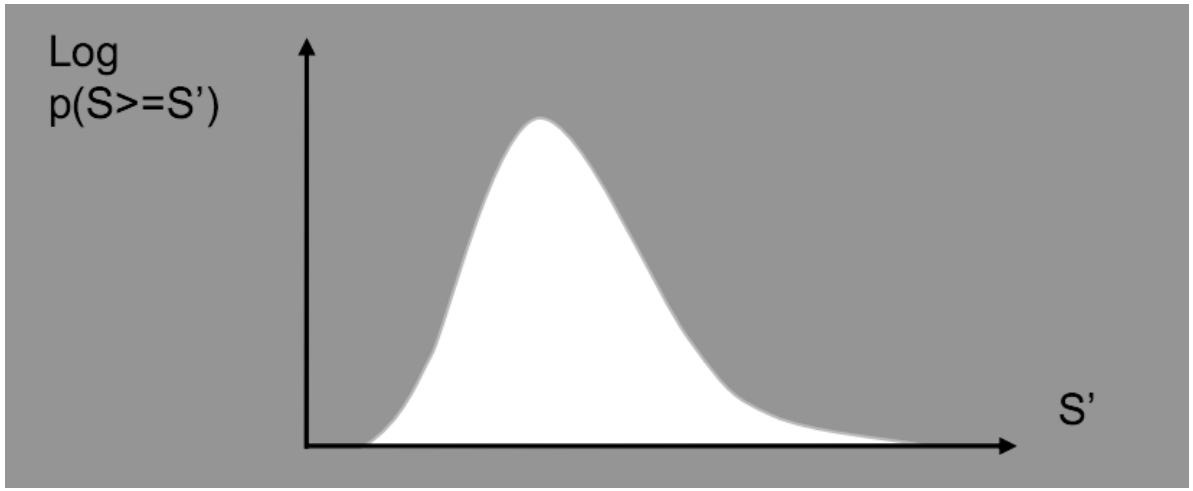
$$E = Kmn \cdot e^{-\lambda S}$$

Istotność statystyczna dopasowania lokalnego

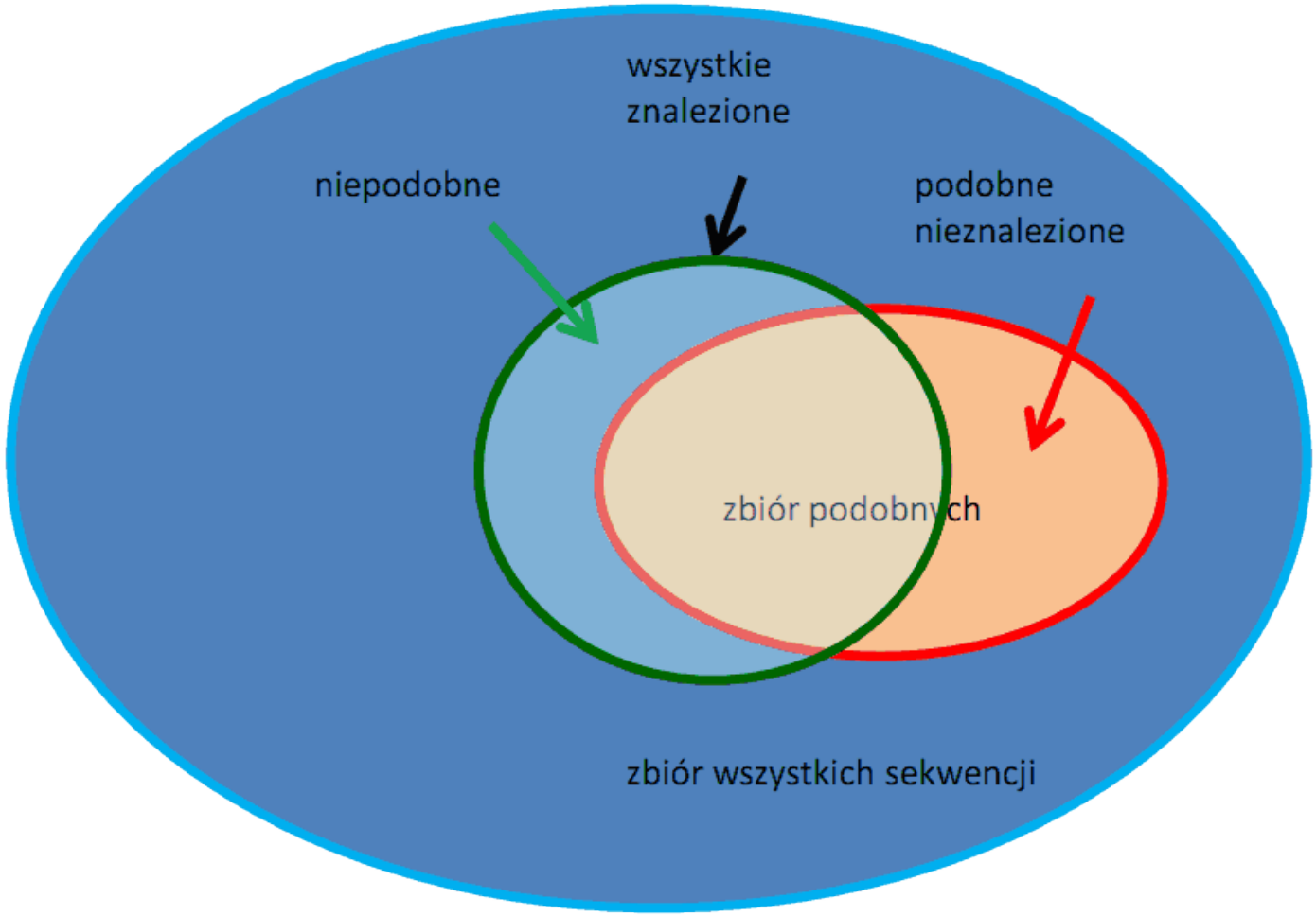
E-value – oczekiwana (wg rozkładu *evd*) liczba dopasowań z punktacją równą przynajmniej S.

Im niższe E, tym bardziej realna homologia.

$$E = Kmn \cdot e^{-\lambda S}$$



Nasza wiedza wobec wszystkich istniejących sekwencji.



Metody heurystyczne

Zadanie:
Porównać badaną sekwencję ze wszystkimi sekwencjami w bazie danych.

Problem:
- Liczba danych jest ogromna i wciąż rośnie
  np. baza nukleotydowa: 300 000 000 000 nukleotydów

- Czas pracy komputera z mocą obliczeniową 10^7 komórek macierzy na sekundę (pełne zestawienie metodami dynamicznymi):
  – białko 300 aminokwasów: ponad 24 godziny dla bazy białkowej
  – DNA 1000 nukleotydów: ponad 300 dni dla bazy w GenBanku

Metody heurystyczne - idea

Programowanie dynamiczne zapewnia najlepsze zestawienie,
jest jednak czasochłonne

ale

większość sekwencji w bazie danych nie jest homologiczna
do sekwencji badanej

zatem

znalezienie sposobu na ich odrzucenie przyspieszyłoby obliczenia!

Strategia

– szybkie przejrzenie bazy sekwencji

– wyeliminowanie sekwencji niepodobnych

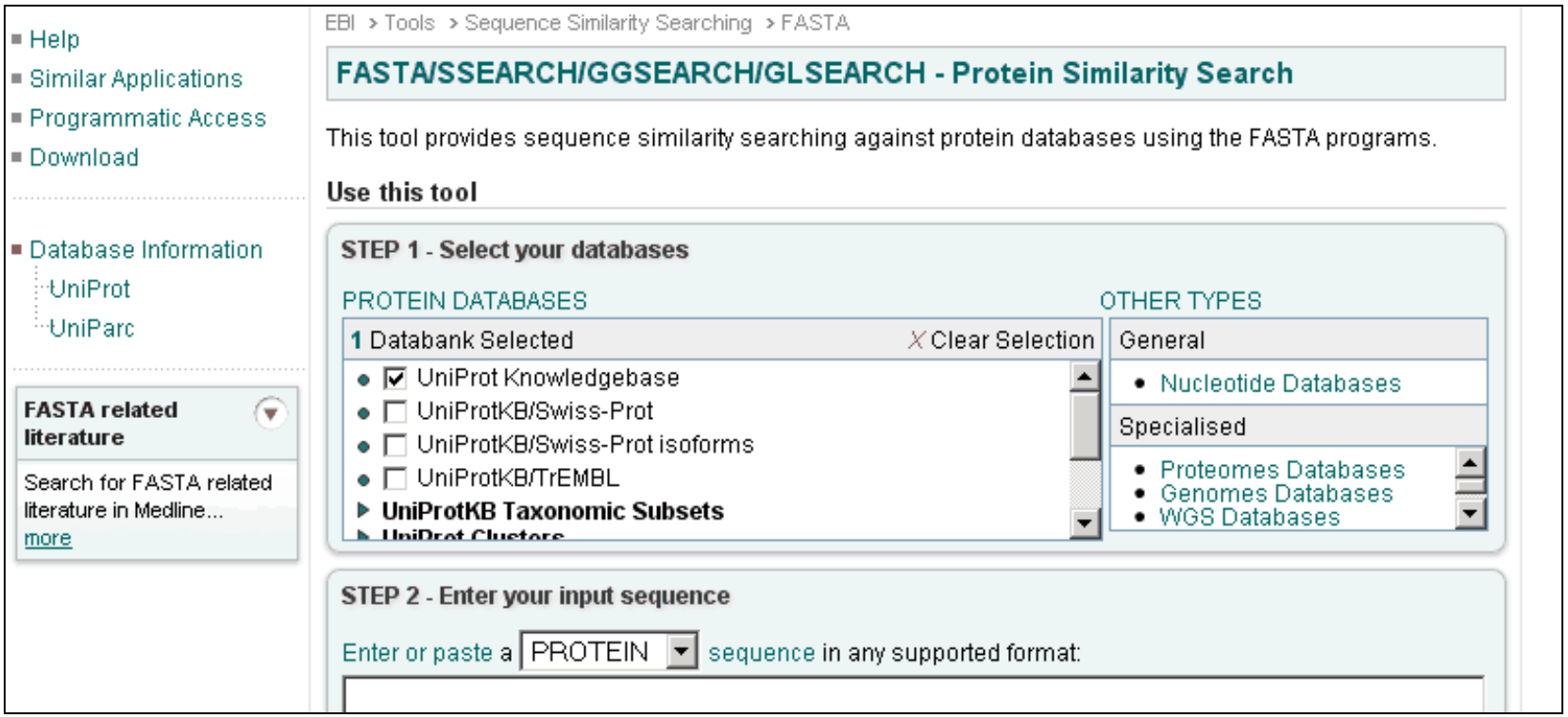
– zestawienie *(alignment)* najlepszych

Najczęściej stosowane algorytmy to **BLAST** i **FASTA**.

BLAST i FASTA - porównanie

BLAST	FASTA
może podawać więcej niż jeden region o wysokiej punktacji	podaje tylko jedno najlepsze dopasowanie
lepszy dla sekwencji białek niż DNA	lepszy dla sekwencji DNA niż białek
szybszy niż FASTA	wolniejszy niż BLAST
mniej czuły niż FASTA przy użyciu domyślnych ustawień	bardziej czuły niż BLAST
daje gorsze rozróżnienie między prawdziwymi i fałszywymi homologami	daje lepsze rozróżnienie między prawdziwymi i fałszywymi homologami

Interfejs www:
http://www.ebi.ac.uk/Tools/sss/fasta/



Publikacje:
- Lipman DJ, Pearson WR. ”Rapid and sensitive protein similarity searches.”Science.(1985)227(4693):1435-41.
- Pearson WR, Lipman DJ. ”Improved tools for biological sequence comparison.” Proc Natl Acad Sci USA 1988)85(8):2444-8.

Algorytm BLAST

BLAST = *Basic Local Alignment Search Tool*

Oparty na wynikach statystycznego rozkładu punktacji lokalnych zestawień.

Zasada działania podobna do FASTA:

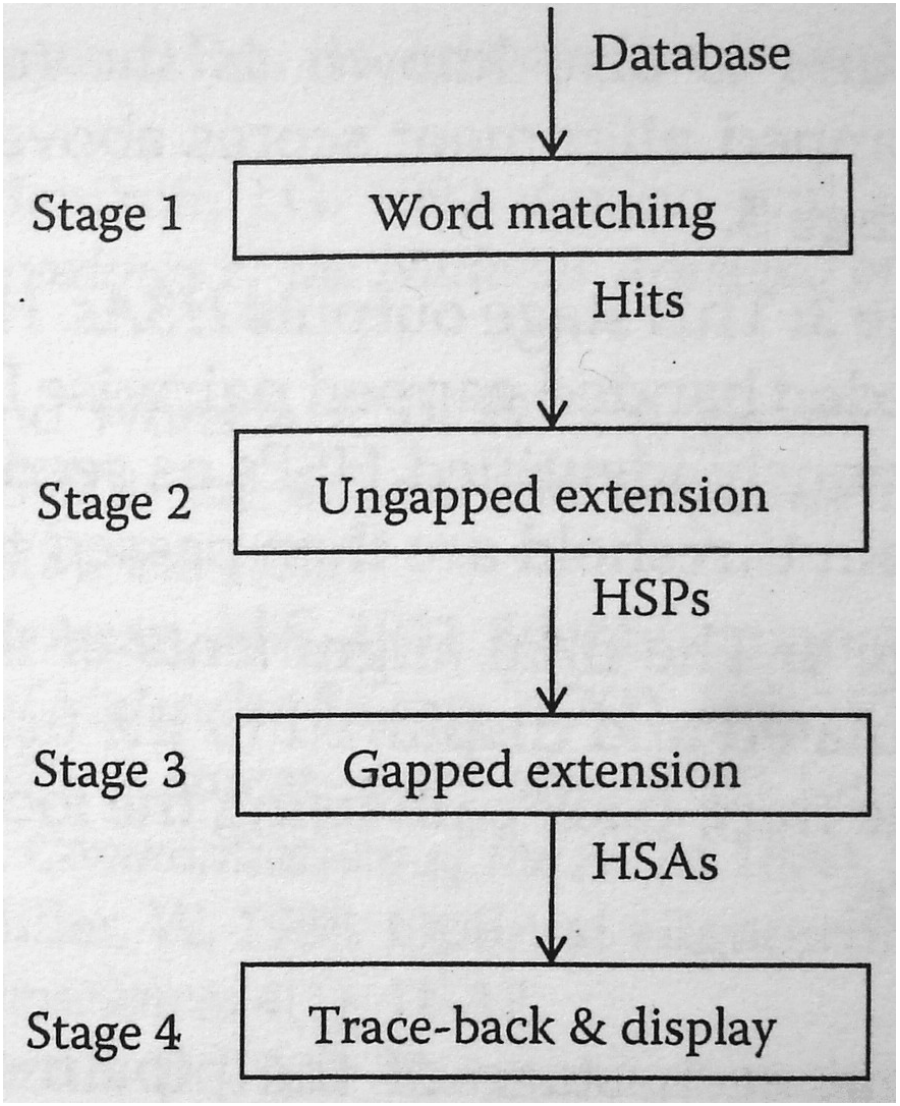
- szybkie przeszukanie bazy, odrzucenie niepodobnych sekwencji, formalne zestawienie;
- rozkład punktacji zestawień bez przerw (gaps) można wyliczyć; zestawienia z przerwami tworzy się z szeregu zestawień bez przerw.

• *Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.*

• *Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." NAR 25:3389-3402.*

Idea filtracji

Zakładamy, ze dobre dopasowanie zawiera krótkie fragmenty z pełną zgodnością.

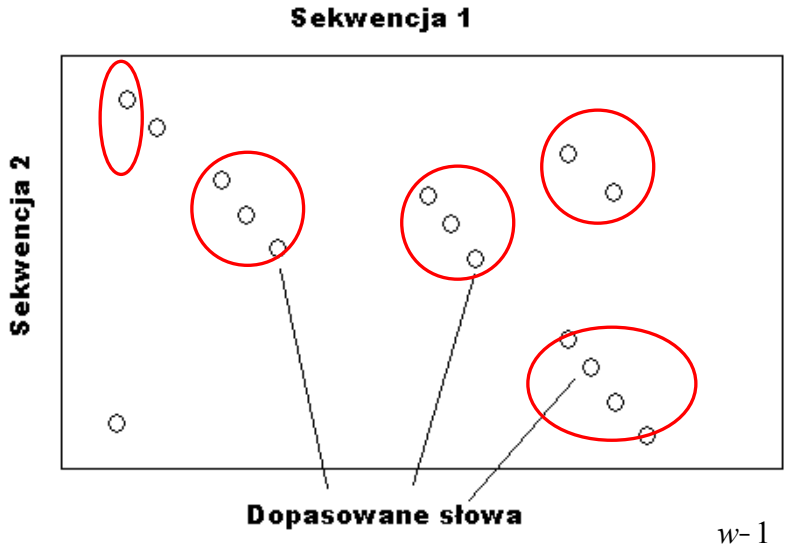


Dopasowanie słów

Heurystyka dopasowania słowa:

Dobre alignmenty zawierają krótkie bardzo podobne obsza
w obu sekwencjach (np. motywy związane z katalizą)

BLAST tworzy listę krótkich słów występujących w danej sekwencji i ich najbliższych sąsiadów.



$$\sum_{k=0}^{w-1} sbt(Q[i+k], D[j+k]) \geq T$$

W=3

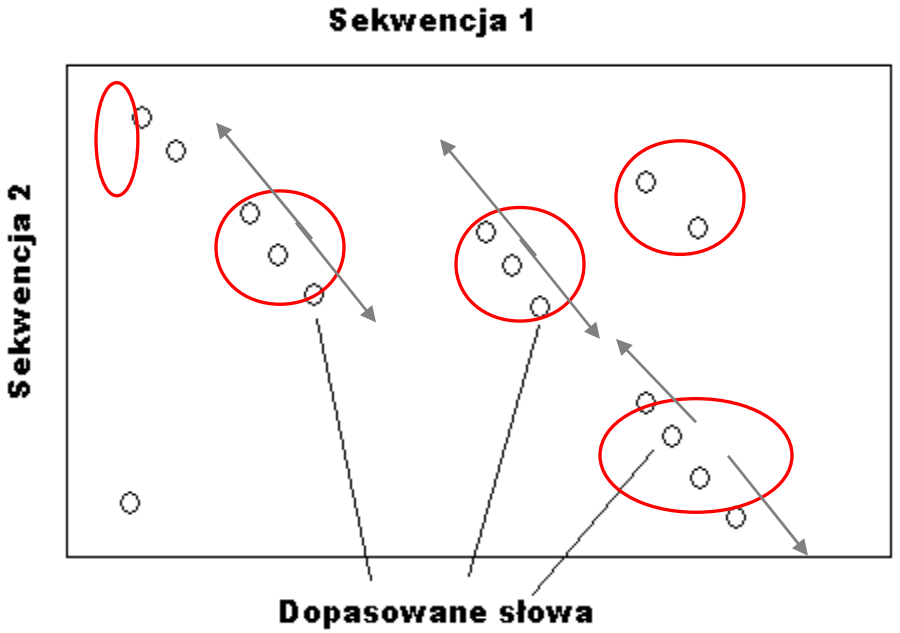
T=12

RGD	17
KGD	14
QGD	13
RGE	13
EGD	12
HGD	12
NGD	12
RGN	12
AGD	11
MGD	11
RAD	11
RGQ	11
RGS	11
RND	11
RSD	11
SGD	11
TGD	11

HSP – *High-scoring Segment Pairs*

Identyfikujemy pozbawione przerw rozszerzenie na diagonali ***d***, zawierające nienakładające się trafienia par (i_1,j_1), (i_2,j_2) wewnątrz okna ***A***, tzn.

$d=i_1-j_1=i_2-j_2,\ w \le i_2-i_1 \le A$



Identyfikacja HSA

Dopasowanie (programowanie dynamiczne) na bazie wcześniej zidentyfikowanych HSPs.

Dopasowania o wartości przekraczającej zadany próg przechodzą do kolejnego etapu.

→ Obliczamy finalne dopasowania dla kompletnych, najwyżej punktowanych sekwencji.

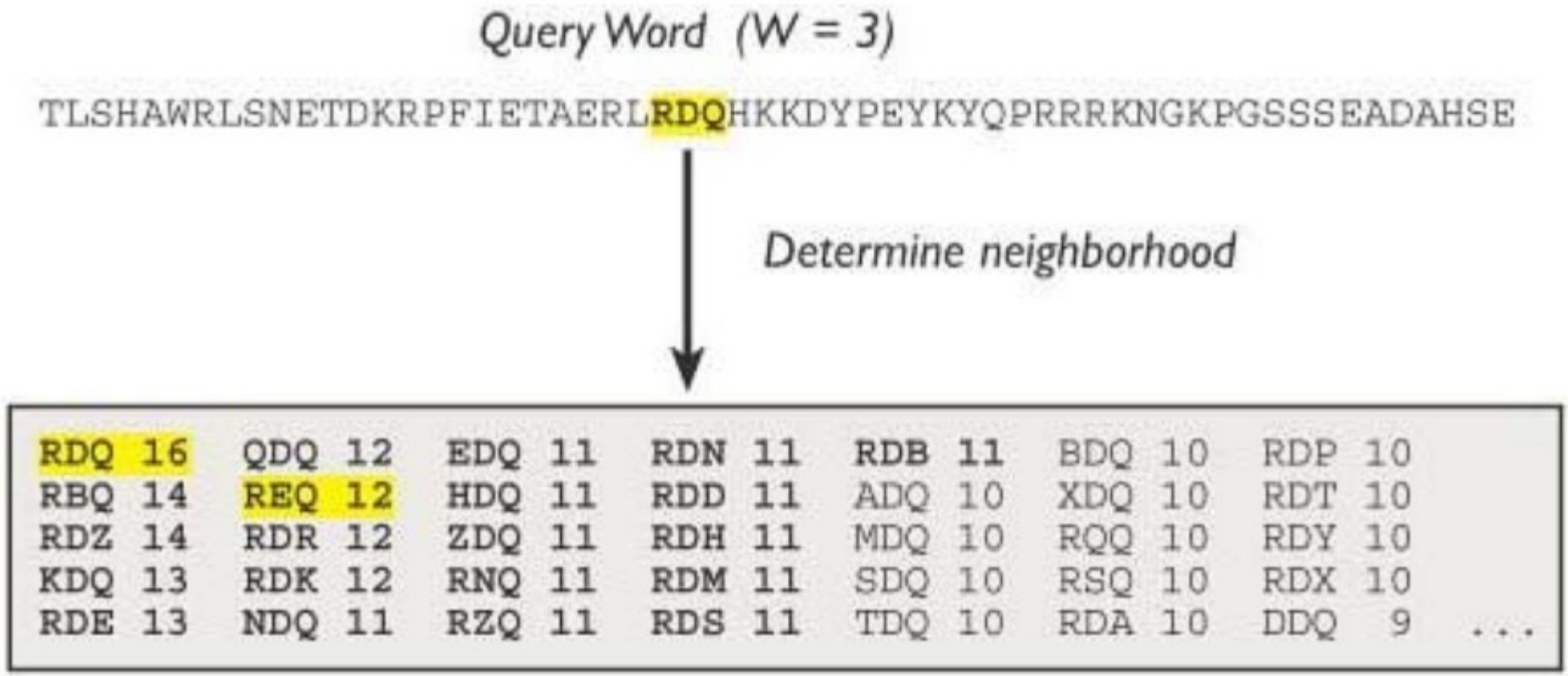
Znaczenie parametrów

ze wzrostem **długości słowa w** – wzrasta szybkość i specyficzność, maleje czułość porównań (większe ziarna)

ze wzrostem **progu dla podobnych słów T** – wzrasta szybkość i specyficzność, spada czułość porównań (mniej ziaren)

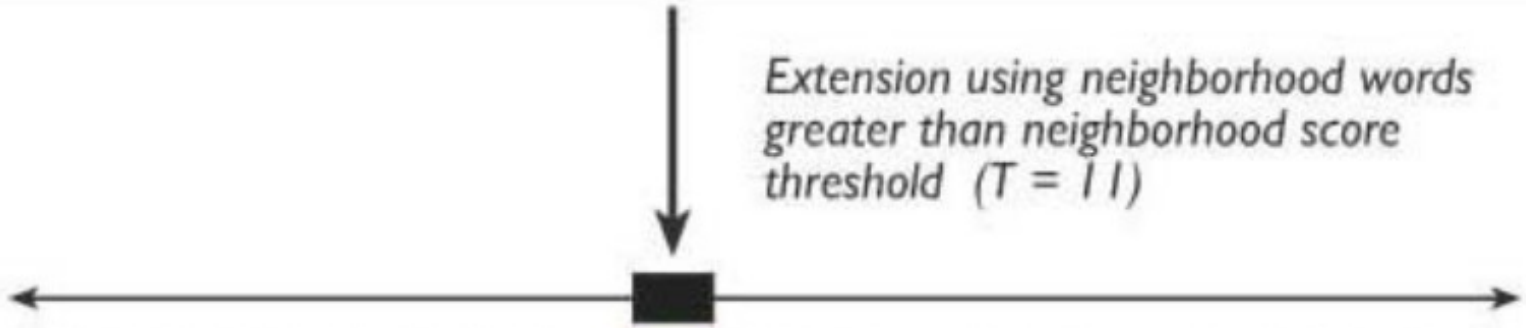
ze wzrostem **progu (odcięcia) dla rozszerzania alignmentu X** – wzrasta szybkość (zmniejsza się długość alignmentów),
zmniejsza się czułość, wzrasta specyficzność (nie są zawierane przypadkowe obszary)

Przykład (1) – dopasowanie podobnych słów



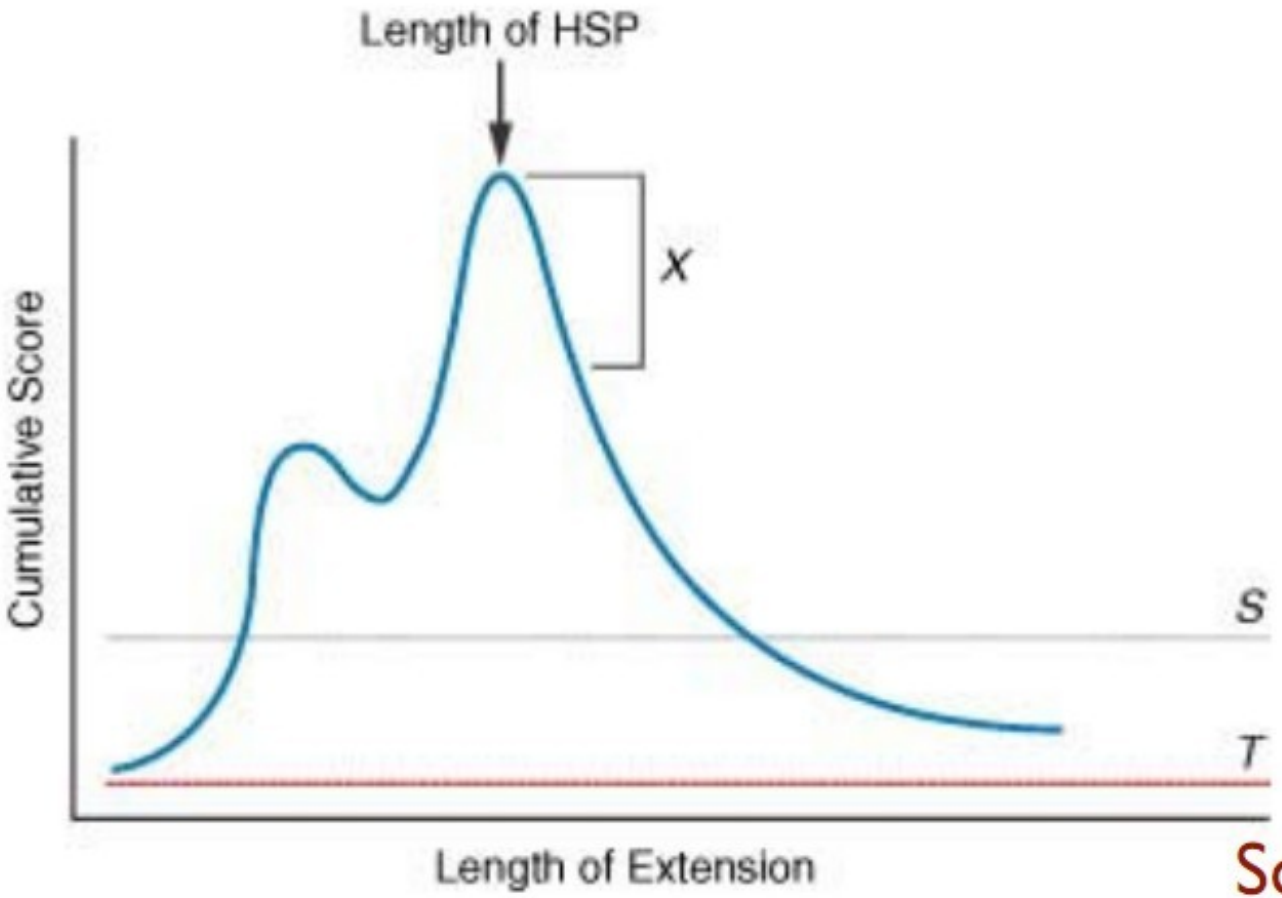
Przykład (2) – rozszerzenie dopasowania

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...




```
Query:  1   TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE  58
            TL   WRL N  +KRPF+E AERLR+QHKKD+P+YKYQPRRRK+ K G S   D   +
Sbjct: 140 TLESGWRLENPGEKRPFVEGAERLREQHKKDHPDYKYQPRRRKSVKNGQSEPEDGSEQ 197
```

Przykład (3) – określenie HSP



Minimum Score (S)

Neighborhood Score Threshold (T)

27

Artykuł

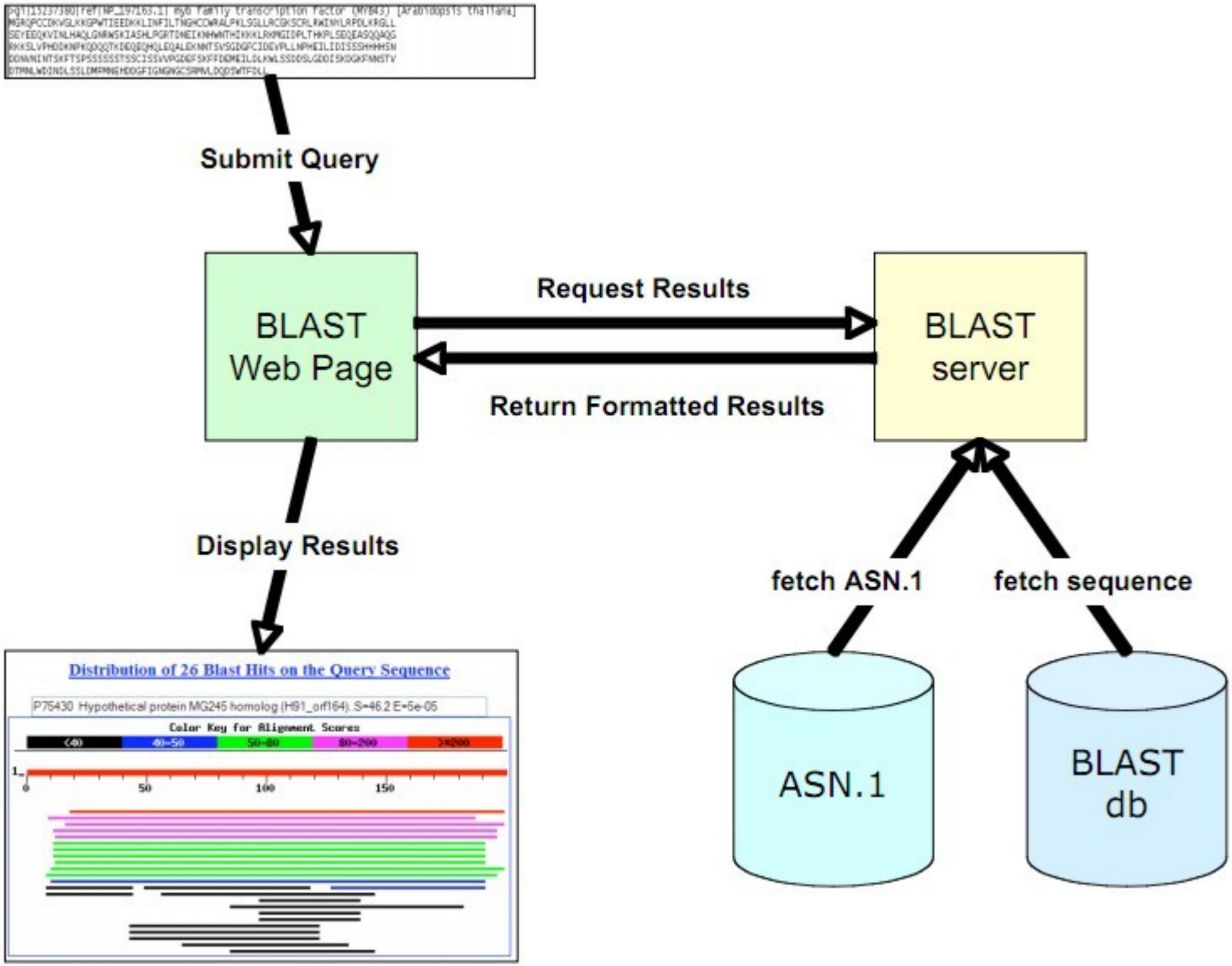
Oryginalny artykuł z 1990 roku:

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman: **Basic Local Alignment Search Tool**

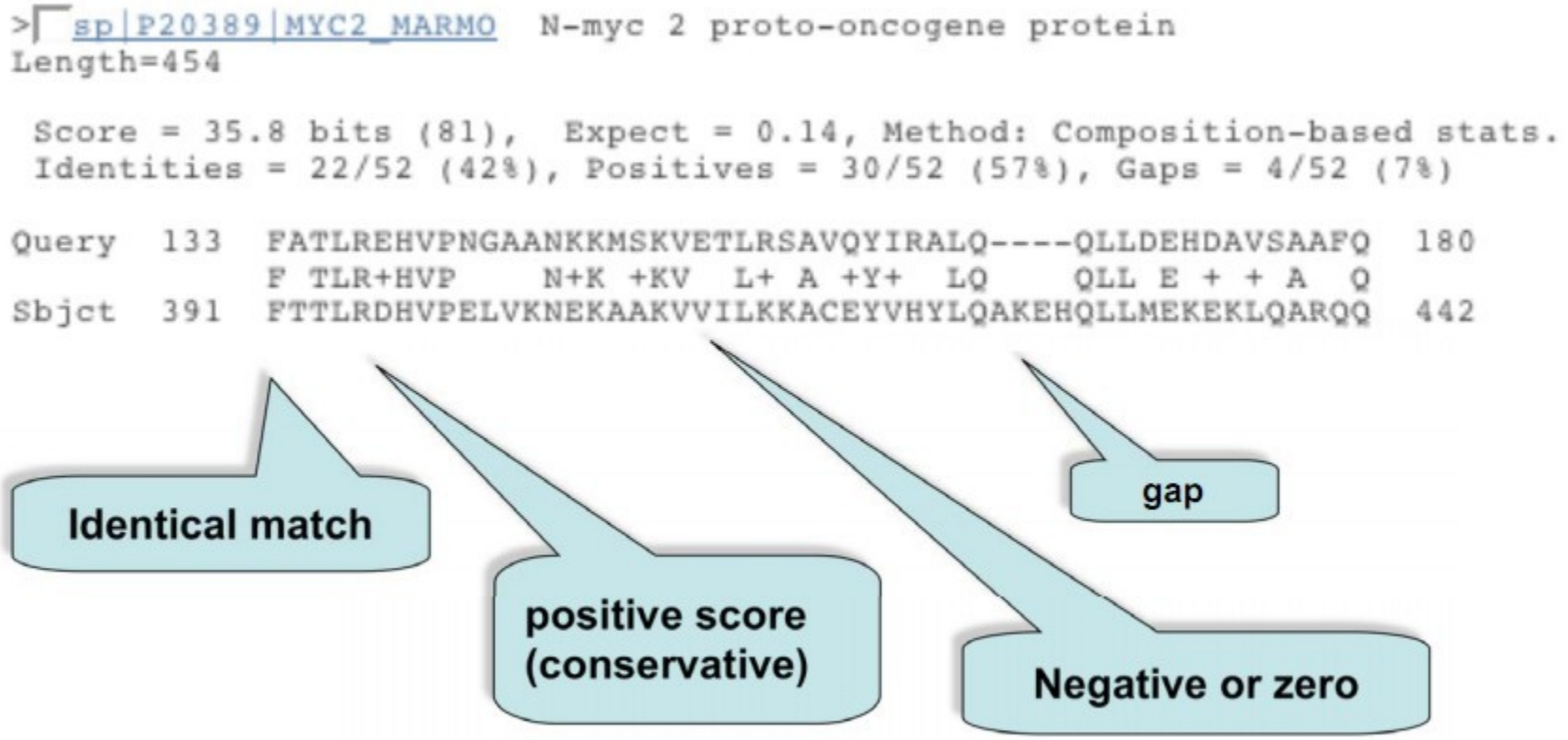
http://www.blastalgorithm.com/

Jest to jeden z najczęściej cytowanych artykułów naukowych.

Aplikacja BLAST



Interpretacja wyników



BLAST w praktyce

Rodzina BLAST

Program	Baza	Sekwencja	Typowe użycie
BLASTN	nt	nt	Mapowanie oligonukleotydów, sekwencji EST, powtórzeń, identyfikacja pokrewnych transkryptów.
BLASTP	białko	białko	Identyfikacja wspólnych rejonów między białkami. Zbieranie białek do analizy filogenetycznej.
BLASTX	białko	nt	Szukanie kodujących sekwencji w genomach.
TBLASTN	nt	białko	Identyfikowanie transkryptów potencjalnie kodujących pokrewne białka (białka jeszcze nie w GenBanku). Mapowanie białek do genomu.
TBLASTX	nt	nt	Przewidywanie genów na podstawie ortologów (z innego gatunku). Poszukiwanie genów „gubionych” przez tradycyjne metody.

Narzędzia (strona EBI)

Sequence Similarity Searching

http://www.ebi.ac.uk/Tools/sss/

Tools > Sequence Similarity Searching

Sequence Similarity Searching is a method of searching sequence databases by using alignment to a query sequence. By statistically assessing how well database and query sequences match one can infer homology and transfer information to the query sequence.

The tools can be launched with different form pre-sets using the links - these can be changed on the tool page as well.

FASTA

FASTA

FASTA is another commonly used sequence similarity search tool which uses heuristics for fast **local** alignment searching.

Protein Nucleotide Genomes Whole Genome Shotgun

SSEARCH

SSEARCH is an optimal (as opposed to heuristics-based) **local** alignment search tool using the Smith-Waterman algorithm. Optimal searches guarantee you find the best alignment score for your given parameters.

Protein Nucleotide Genomes Whole Genome Shotgun

PSI-Search

PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST profile construction strategy to find distantly related protein sequences.

Protein

GGSEARCH

GGSEARCH performs optimal **global-global** alignment searches using the Needleman-Wunsch algorithm.

Protein Nucleotide

GLSEARCH

GLSEARCH performs an optimal sequence search using alignments that are **global**

BLAST

NCBI BLAST

NCBI BLAST (blastall) is the most commonly used sequence similarity search tool. It uses heuristics to perform fast **local** alignment searches.

Protein Nucleotide Vectors

WU-BLAST

WU-BLAST is similar to NCBI BLAST but combines multiple parameter options into a simpler 'sensitivity' setting.

Protein Nucleotide

PSI-BLAST

PSI-BLAST allows users to construct and perform a BLAST search with a custom, position-specific, scoring matrix which can help find distant evolutionary relationships. PHI-BLAST functionality is also available to restrict results using patterns.

Protein

ENA Sequence Search

The EBI has a new search tool which is far faster than BLAST, with only a marginal loss in search sensitivity.

Try it out at ENA Sequence Search.

Biopython – BLAST przez WWW

Podręcznik, rozdział 7:
http://biopython.org/DIST/docs/tutorial/Tutorial.html#htoc87

1)
```
>>> from Bio.Blast import NCBIWWW
>>> help(NCBIWWW.qblast)
```
2)
```
>>> from Bio.Blast import NCBIWWW
>>> result_handle = NCBIWWW.qblast("blastn", "nt", "8332116„)
```
3)
```
>>> from Bio.Blast import NCBIWWW
>>> fasta_string = open("m_cold.fasta").read()
>>> result_handle = NCBIWWW.qblast("blastn", "nt", fasta_string)
```
4)
```
>>> from Bio.Blast import NCBIWWW
>>> from Bio import SeqIO
>>> record = SeqIO.read("m_cold.fasta", format="fasta")
>>> result_handle = NCBIWWW.qblast("blastn", "nt", record.seq)
```

BLAST lokalnie

Zalety:
- szybszy
- możliwość przechowywania dedykowanej bazy danych

Narzędzie: BLAST+

```
blastx -query opuntia.fasta -db nr -out opuntia.xml -evalue 0.001 -outfmt 5
```

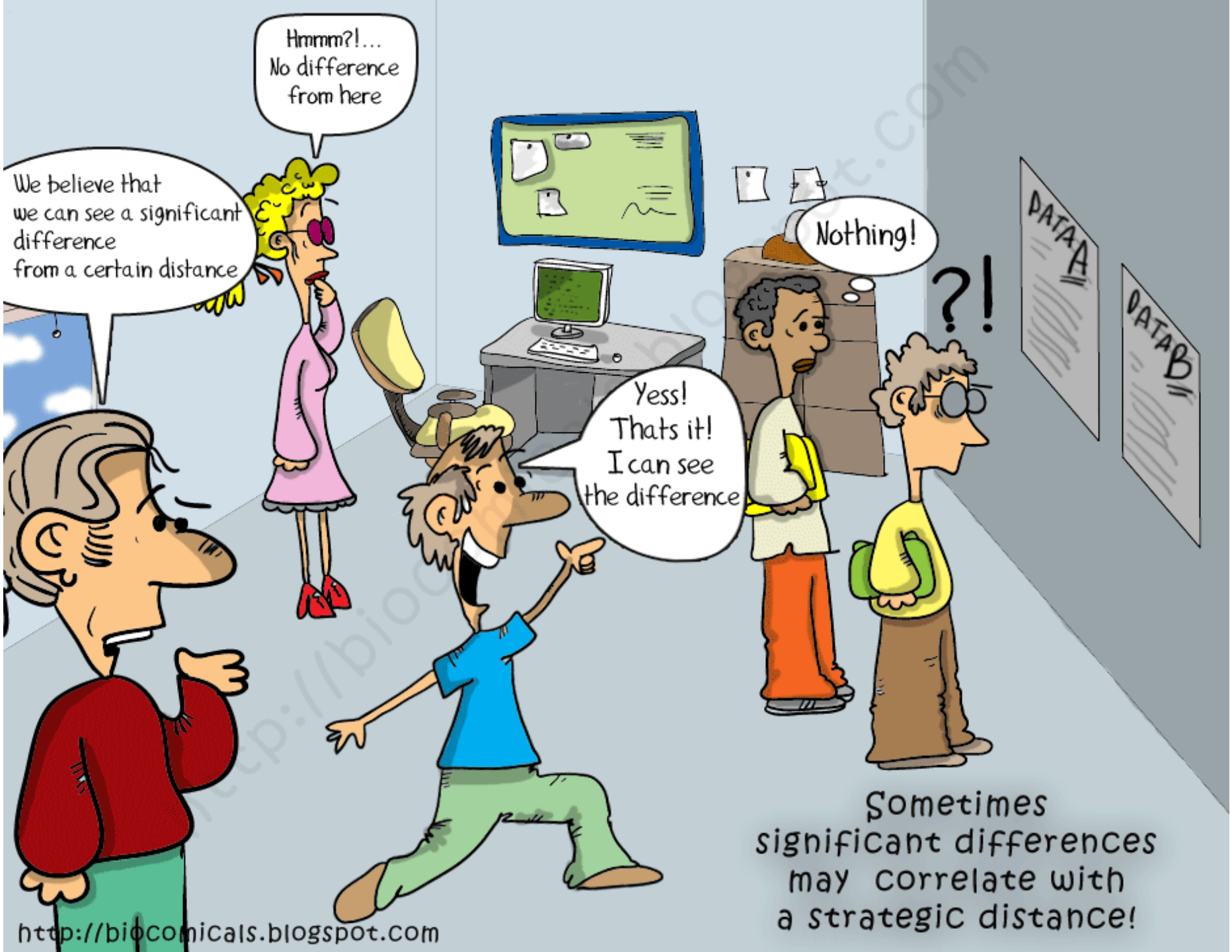
Powyższe polecenie uruchamia algorytm blastx na nieredundantnej bazie danych (nr) z obcięciem miary istotności na poziomie 0,001 i zapisuje wynik do pliku, w formacie XML.

BLAST lokalnie

```
*** Formatting options
 -outfmt <String>
   alignment view options:
     0 = pairwise,
     1 = query-anchored showing identities,
     2 = query-anchored no identities,
     3 = flat query-anchored, show identities,
     4 = flat query-anchored, no identities,
     5 = XML Blast output,
     6 = tabular,
     7 = tabular with comment lines,
     8 = Text ASN.1,
     9 = Binary ASN.1,
    10 = Comma-separated values,
    11 = BLAST archive format (ASN.1)
```

BLAST lokalnie - biopython

```
>>> from Bio.Blast.Applications import NcbiblastxCommandline

>>> help(NcbiblastxCommandline) ...

>>> blastx_cline = NcbiblastxCommandline(query="opuntia.fasta", db="nr",
evalue=0.001, ... outfmt=5, out="opuntia.xml")

>>> blastx_cline NcbiblastxCommandline(cmd='blastx', out='opuntia.xml',
outfmt=5, query='opuntia.fasta', db='nr', evalue=0.001)

>>> print(blastx_cline) blastx -out opuntia.xml -outfmt 5 -query opuntia.fasta
-db nr -evalue 0.001

>>> stdout, stderr = blastx_cline()
```


We believe that we can see a significant difference from a certain distance
Hmmm?!... No difference from here
Yess! Thats it! I can see the difference
Nothing!
?!
DATA A
DATA B
Sometimes significant differences may correlate with a strategic distance!
http://biocomicals.blogspot.com