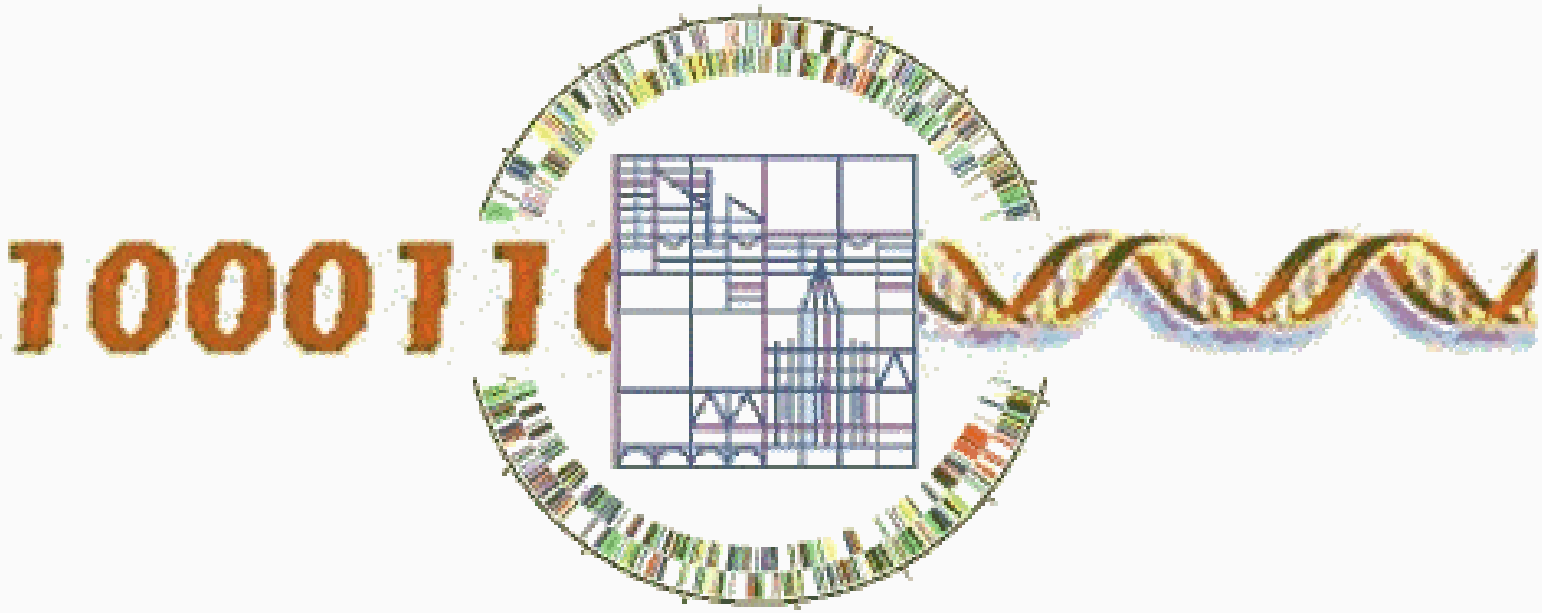
Bioinformatics and Drug Research

Klaus P. Schaefer



Sequence Searching and Analysis II

Summer Course - Lecture 04 - Version of May 26, 2003

http://www.uni-konstanz.de/FuF/Bio/Bioinformatik/



Analysis of Protein Sequences

- **Calculation of physico-chemical Parameters**
  - **Molecular Weight**
  - **Amino Acid Composition**
  - **extinction coefficient**
  - **Isoelectric Point (Ip)**
  - **Charge at pH 7 ... and many more**

- **Online Protein Analysis Program Resources**
  - **these are just 2 examples - there are many more:**
    - ***EXPLORE!***

http://www.up.univ-mrs.fr/~wabim/english/logligne.html

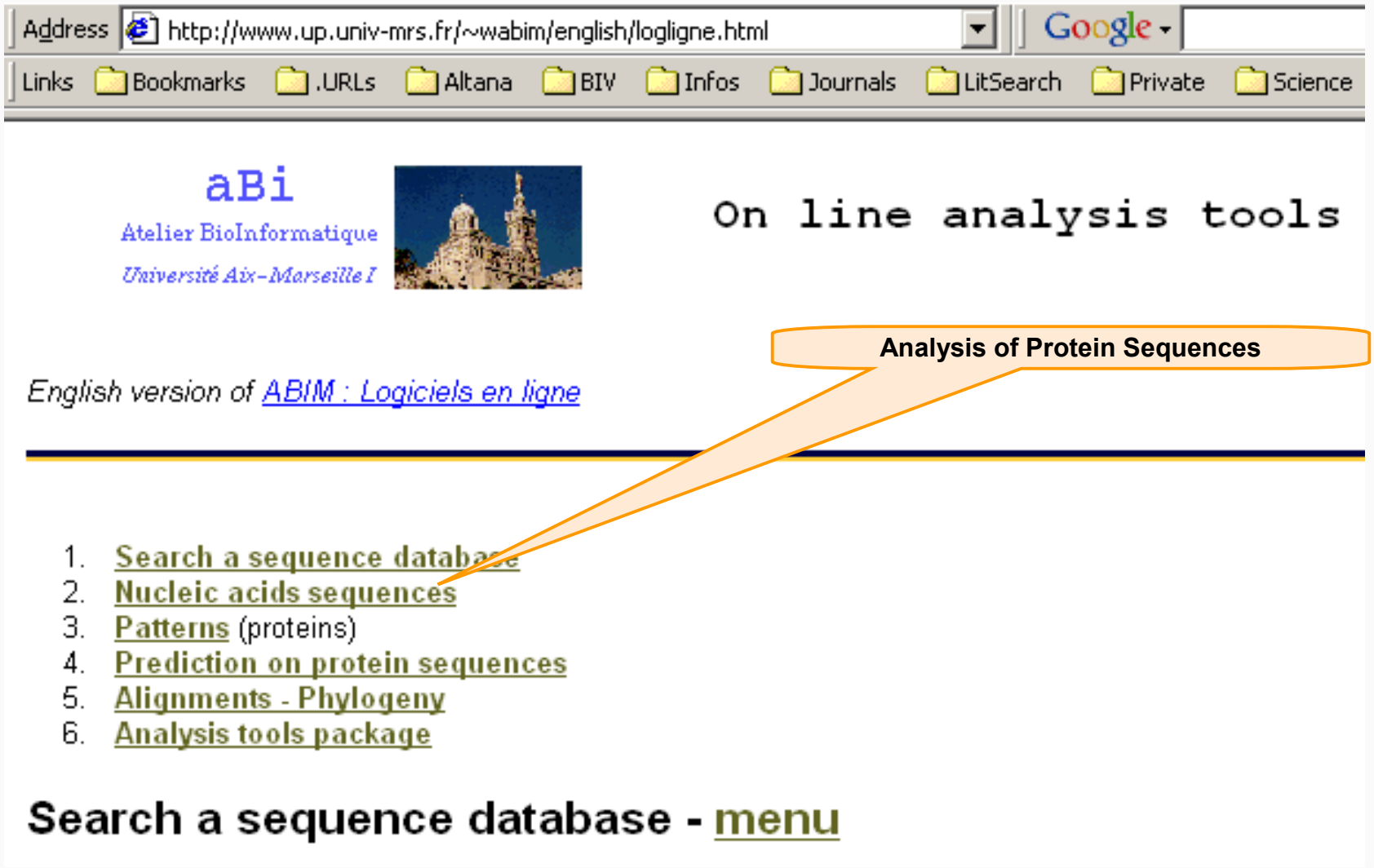
WebSite Infos - On line analysis tools
Search a sequence database, Nucleic acids sequences, Patterns (proteins), Prediction on protein sequences, Alignments – Phylogeny, Analysis tools package

http://www.expasy.ch/tools/

WebSite Infos - ExPASy Proteomics tools
[Protein identification and characterization]
[DNA -> Protein] [Similarity searches]
[Pattern and profile searches]
[Post-translational modification prediction]
[Primary structure analysis]
[Secondary structure prediction] [Tertiary structure]
[Transmembrane regions detection] [Sequence alignment]

An Example for an Online Sequence Analysis Resource

http://www.up.univ-mrs.fr/~wabim/english/logligne.html





This Package offers many Options

Prediction on protein sequences - menu

- Analysis tools package
  - BCM Search Launcher **BCM** (USA)
  - HNB (Family, function, structure) **Berlin** (Germany)
  - META PredictProtein server (SignalP, NetOglyc, NetPicoRNA, ChloroP, JPred, TmHMM, ToPred, DAS, FRSVR, SAMT98, Swiss-Model, CPHmodels) **Columbia** (USA)
  - Network Protein Sequence @nalysis (SOPM, SOPMA, HNN, MLRC, DPM, DSC, GOR, PHD, PREDATOR,..) **PBIL** (France

- Analyze
  - AA Composition, MW, OD, pI, titration curve **ABIM** (in french)
  - AA Composition **SBDS** (UK), see also **BMB**
  - pI et Mw **Expasy**
  - ASAP (Analysis of Sequence and Aminoacid Probabilities) **Rockefeller** (USA)
  - ASC **IMP** (Analysis of protein surface : Van-der-Waals, energy, Hydrophilicity, Hydrophobicity)
  - CAST (Compositional Bias Detection Algorithm) **EBI** (UK)
  - CUTTER (generate peptide fragments by the enzymatic or chemical cleavage) **Tours** (France)
  - CysPred (predicting the redox state of Cys residues from sequences or multiple sequence alignments) **Rockfeller** (USA), see also another **link**, see also **CIRB**
  - EMBOSS at Pasteur (France)
    - Protein charge plot **CHARGE**
    - Protein proteolytic enzyme or reagent cleavage digest **DIGEST**
    - Build domain coordinate files **DOMAINER**
    - Displays a wordmatch dotplot of two sequences **DOTTUP**
    - Calculates Km and Vmax for an enzyme reaction **FINDKM**
    - Hydrophobic moment calculation **HMOMENT**
    - Displays protein hydropathy **PEPWINDOW**
    - Displays protein hydropathy of a set of sequences **PEPWINDOWALL**
    - Displays protein hydropathy **OCTANOL**
    - Calculates the isoelectric point of a protein **IEP**
    - Protein statistics **PEPSTATS**

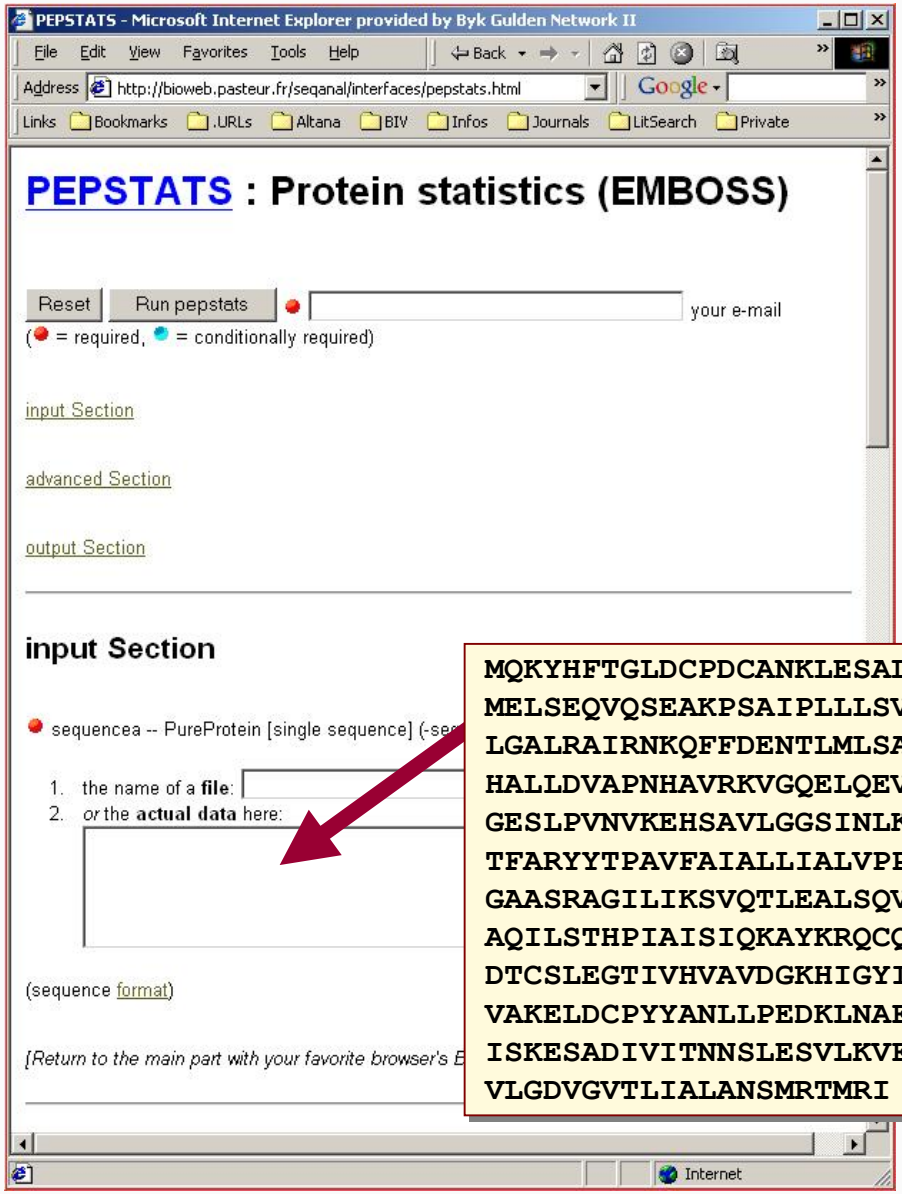
Selecting PEPSTATS



Analyzing three quite distinct Proteins: SP-A, SP-C & CadA

- **SP-A** *P07714*
  - is a glycoprotein in the lung surfactant of vertebrates
  - it is a lectin
  - it is an opsonin which binds to the surface of microorganisms and viruses tagging them for recognition by alveolar macrophages
  - it has a collagen-like domain
- **SP-C** *NM_003018*
  - highly hydrophobic protein of 34/35 amino acids
  - synthesized as a 192 aa precursor protein
  - integrates into phospholipid matrix and regulates surface tension
- **CadA** *AF125316*
  - is a transmembrane ion pump in bacteria
  - it is a P type ATPase
  - it exports transition metal ions, e.g. Cd2+, Zn2+, Ni2+, Co2+ etc.
  - by this action it protects the cell against the toxic effects of these metal ions

Cut and Paste your sequence into the Input Window




```
MQKYHFTGLDCPDCANKLESALNKQSYVQEARVVFNTNTLYLKTQDMPKALALIKQLEPD
MELSEQVQSEAKPSAIPLLLSVVLYLIAVATIHFSAQNWALHLSYALLAGVYLVAGKDVF
LGALRAIRNKQFFDENTLMLSATIAAFGVGAHEEAVSIMVFYSAGEFLQQLAIARSKQSL
HALLDVAPNHAVRKVGQELQEVEPSALAIGDVVIVKVGEKIPTDGVVLRGESLLDQKALT
GESLPVNVKEHSAVLGGSINLKGVLEIQVSKLYADSSVAKIVDLVSNAVSQKSQTEKFIT
TFARYYTPAVFAIALLIALVPPLLGHGDFDTWIYRGLFALMVSCPCALVISVPLGYFGGV
GAASRAGILIKSVQTLEALSQVKNIAFDKTGTLSKGEFNVIDVVPIAPFSKEDVLQHATC
AQILSTHPIAISIQKAYKRQCQHEVKDYQEIGGLGVQASCHSHLIIAGNDKMLHKYNIPH
DTCSLEGTIVHVAVDGKHIGYIVVADTLKDNAKECLDGLKHAGIEHMCILSGDHEYSTKR
VAKELDCPYYANLLPEDKLNAFKDFQAQHAHKSMFVGDGINDAPTLARADVSMSMGSASQ
ISKESADIVITNNSLESVLKVFKIAKKTKRIIIENIIFALAIKAMFIVLGLSGDASLWEA
VLGDVGVTLIALANSMRTMRI
```

SP-A is a Member of the Lung Surfactant Proteins



```
PEPSTATS of SP-A, from 1 to 248
Molecular weight = 26255.47             Residues = 248
Average Residue Weight  = 105.869       Charge   = -6.0
Isoelectric Point = 4.8177
Residue          Number          Mole%           DayhoffStat
A = Ala          18              7.258           0.844
B = Asx          0               0.000           0.000
C = Cys          9               3.629           1.251
D = Asp          12              4.839           0.880
E = Glu          17              6.855           1.142
F = Phe          6               2.419           0.67
G = Gly          38              15.323          1.824
H = His          4               1.613           0.806
I = Ile          10              4.032           0.896
K = Lys          7               2.823           0.4
L = Leu          18              7.258           1
M = Met          6               2.419           1.423
N = Asn          10              4.032           0.938
P = Pro          26              10.484          2.016
Q = Gln          8               3.226           0.827
R = Arg          14              5.645           1.152
S = Ser          11              4.435           0.634
T = Thr          11              4.435           0.727
V = Val          12              4.839           0.733
W = Trp          3               1.210           0.931
X = Xxx          0               0.000           0.000
Y = Tyr          8               3.226           0.949
Z = Glx          0               0.000           0.000
```

Glycine and Proline are abundant in the collagenous domain

CadA of *Helicobacter felis* is an Ion Export Pump



```
PEPSTATS of Hfelis CadA, from 1 to 681
Molecular weight = 73747.36             Residues = 681
Average Residue Weight  = 108.293       Charge   = 4.0
Isoelectric Point = 6.8175
Residue          Number          Mole%           DayhoffStat
A = Ala          75              11.013          1.281
B = Asx          0               0.000           0.000
C = Cys          11              1.615           0.557
D = Asp          34              4.993           0.908
E = Glu          34              4.993           0.832
F = Phe          24              3.524           0.979
G = Gly          43              6.314           0.752
H = His          22              3.231           1.615
I = Ile          51              7.489           1.664
K = Lys          46              6.755           1.0
L = Leu          77              11.307          1.528
M = Met          14              2.056           1.209
N = Asn          23              3.377           0.785
P = Pro          21              3.084           0.593
Q = Gln          30              4.405           1.130
R = Arg          15              2.203           0.450
S = Ser          50              7.342           1.049
T = Thr          28              4.112           0.674
V = Val          61              8.957           1.357
W = Trp          3               0.441           0.339
X = Xxx          0               0.000           0.000
Y = Tyr          19              2.790           0.821
Z = Glx          0               0.000           0.000
```

Histidine is much more abundant than in <average> proteins because it is part of transition metal binding sites



Simple Statistics reveals Important Differences

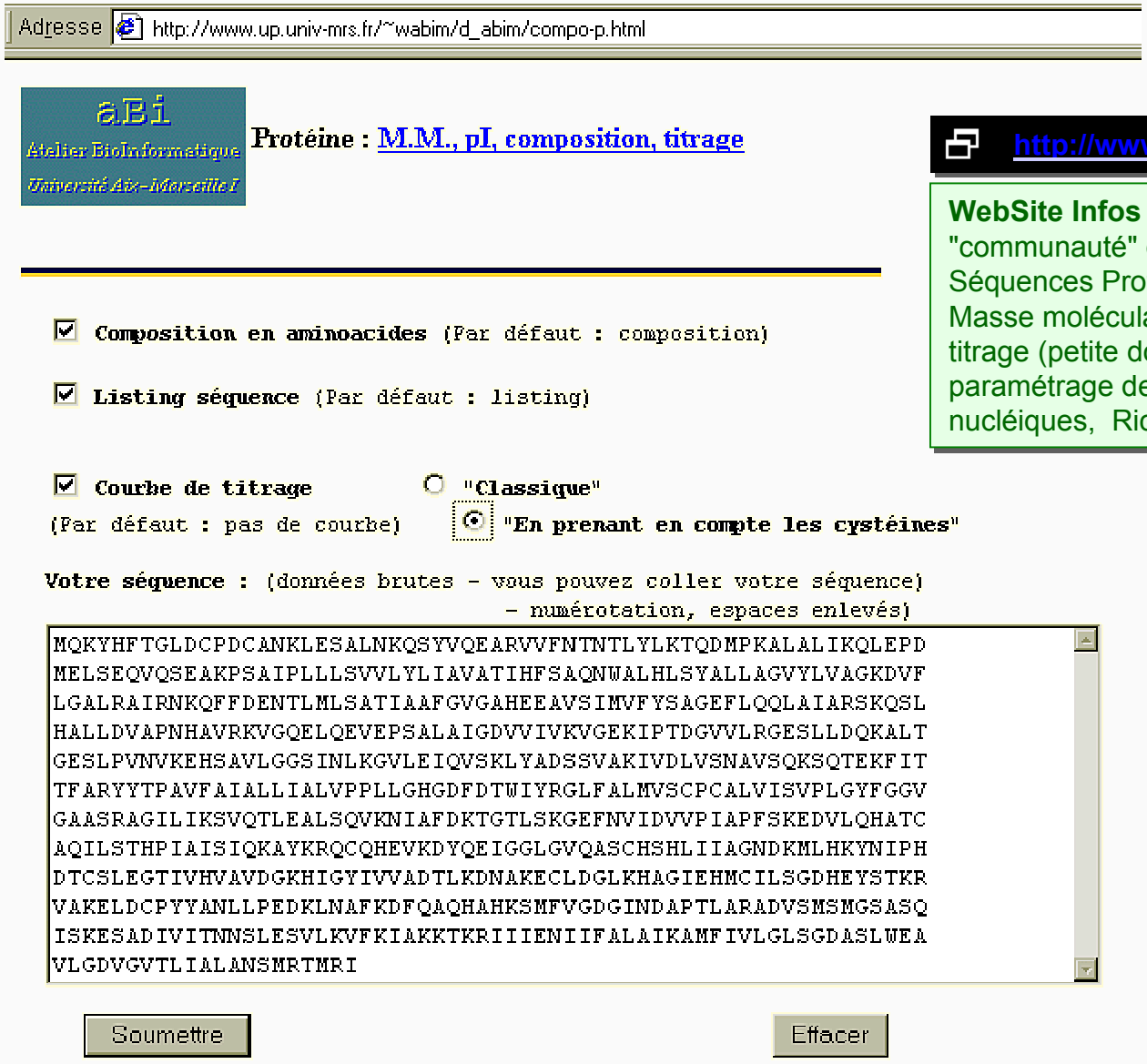
SP-A

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	87	35.081
Small	(A+B+C+D+G+N+P+S+T+V)	147	59.274
Aliphatic	(I+L+V)	40	16.129
Aromatic	(F+H+W+Y)	21	8.468
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	154	62.097
Polar	(D+E+H+K+N+Q+R+S+T+Z)	94	37.903
Charged	(B+D+E+H+K+R+Z)	54	21.774
Basic	(H+K+R)	25	10.081
Acidic	(B+D+E+Z)	29	11.694

CadA

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	207	30.396
Small	(A+B+C+D+G+N+P+S+T+V)	346	50.808
Aliphatic	(I+L+V)	189	27.753
Aromatic	(F+H+W+Y)	68	9.985
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	399	58.590
Polar	(D+E+H+K+N+Q+R+S+T+Z)	282	41.410
Charged	(B+D+E+H+K+R+Z)	151	22.173
Basic	(H+K+R)	83	12.188
Acidic	(B+D+E+Z)	68	9.985

Learning French through Protein Analysis ...



http://www.up.univ-mrs.fr/~wabim/d_abim/

WebSite Infos - Des "petits" programmes pour la "communauté" des biologistes moléculaires : Séquences Protéiques, Riche en .. (petite doc), Masse moléculaire, pI, composition, courbe de titrage (petite doc), Courbe de titrage (petite doc) - paramétrage des valeurs des pK, Séquences nucléiques, Riche en .. (petite doc)

Calculating the Isoelectric Point

```
Protéine : M.M., pI, composition, titrage

Longueur de la séquence :     681 amino acides
Masse moléculaire :     73747,57

Coefficient d'extinction molaire à 280 nm (1cm) :  41390
  - en prenant en compte toutes les cystéines :  42050

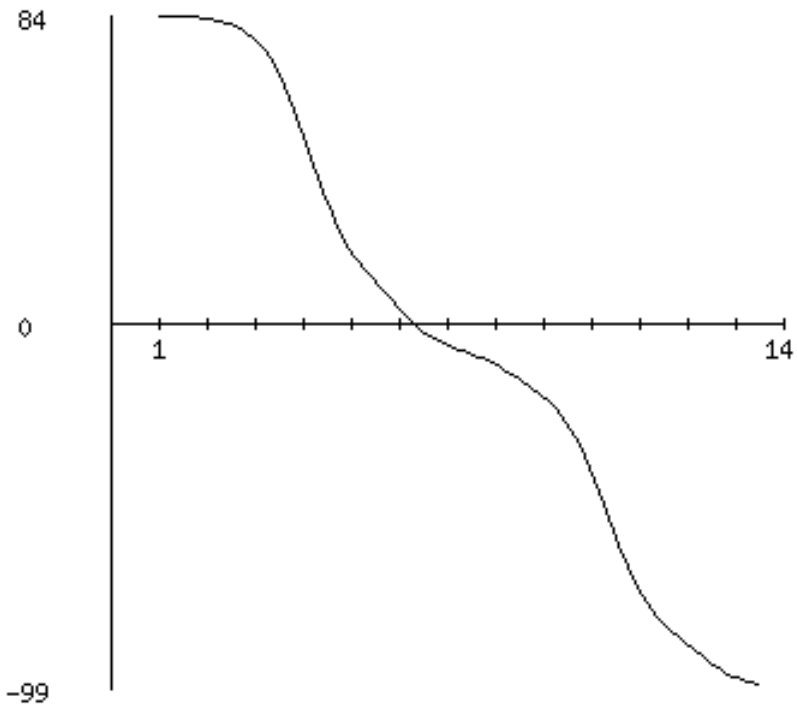

Absorption 0,1% (1g/l) à 280 nm (1cm) :    0,561
  - en prenant en compte toutes les cystéines :    0,570


Point isoélectrique théorique :
   - classique :    6,32
   - en prenant en compte toutes les cystéines :     6,32

Index aliphatique :     110,29


MQKYHFTGLD CPDCANKLES ALNKQSYVQE ARVVFNTNTL YLKTQDMPKA
LALIKQLEPD MELSEQVQSE AKPSAIPLLL SVVLYLIAVA TIHFSAQNWA
LHLSYALLAG VYLVAGKDVF LGALRAIRNK QFFDENTLML SATIAAFGVG
AHEEAVSIMV FYSAGEFLQQ LAIARSKQSL HALLDVAPNH AVRKVGQELQ
EVEPSALAIG DVVIVKVGEK IPTDGVVLRG ESLLDQKALT GESLPVNVKE
HSAVLGGSIN LKGVLEIQVS KLYADSSVAK IVDLVSNAVS QKSQTEKFIT
TFARYYTPAV FAIALLIALV PPLLGHGDFD TWIYRGLFAL MVSCPCALVI
SVPLGYFGGV GAASRAGILI KSVQTLEALS QVKNIAFDKT GTLSKGEFNV
IDVVPIAPFS KEDVLQHATC AQILSTHPIA ISIQKAYKRQ CQHEVKDYQE
IGGLGVQASC HSHLIIAGND KMLHKYNIPH DTCSLEGTIV HVAVDGKHIG
YIVVADTLKD NAKECLDGLK HAGIEHMCIL SGDHEYSTKR VAKELDCPYY
ANLLPEDKLN AFKDFQAQHA HKSMFVGDGI NDAPTLARAD VSMSMGSASQ
ISKESADIVI TNNSLESVLK VFKIAKKTKR IIIENIIFAL AIKAMFIVLG
LSGDASLWEA VLGDVGVTLI ALANSMRTMR I
```

```
Courbe de tirage
Valeurs utilisées des pK :
alpha-NH3 :  9,69        alpha-COOH :  2,34
Arg : 12,40     Lys : 10,50     His :  6,00
Asp :  3,86     Glu :  4,25     Cys :  8,33
Tyr : 10,00
```



Courbe de titrage
(en tenant compte des cysteines)

Note

Analyze the Distribution Pattern of certain Amino Acids

aBi
Atelier BioInformatique
Université Aix-Marseille I

Graphe : protéine riche en :

Ensemble prédéfini AVLIPMFW

Ou votre sélection KR (code une lettre)

Largeur de la fenêtre : 5

Listing séquence (Par défaut : pas de listing)

Votre séquence : (données brutes - vous pouvez coller votre séquence - numérotation, espaces enlev&eacutes)

```
MQKYHFTGLDCPDCANKLESALNKQSYVQEARVVFNTNTLYLKTQDMPKALALIKQLEPD
MELSEQVQSEAKPSAIPLLLSVVLYLIAVATIHFSAQNWALHLSYALLAGVYLVAGKDVF
LGALRAIRNKQFFDENTLMLSATIAAFGVGAHEEAVSIMVFYSAGEFLQQLAIARSKQSL
HALLDVAPNHAVRKVGQELQEVEPSALAIGDVVIVKVGEKIPTDGVVLRGESLLDQKALT
GESLPVNVKEHSAVLGGSINLKGVLEIQVSKLYADSSVAKIVDLVSNAVSQKSQTEKFIT
TFARYYTPAVFAIALLIALVPPLLGHGDFDTWIYRGLFALMVSCPCALVISVPLGYFGGV
GAASRAGILIKSVQTLEALSQVKNIAFDKTGTLSKGEFNVIDVVPIAPFSKEDVLQHATC
AQILSTHPIAISIQKAYKRQCQHEVKDYQEIGGLGVQASCHSHLIIAGNDKMLHKYNIPH
DTCSLEGTIVHVAVDGKHIGYIVVADTLKDNAKECLDGLKHAGIEHMCILSGDHEYSTKR
VAKELDCPYYANLLPEDKLNAFKDFQAQHAHKSMFVGDGINDAPTLARADVSMSMGSASQ
ISKESADIVITNNSLESVLKVFKIAKKTKRIIIENIIFALAIKAMFIVLGLSGDASLWEA
VLGDVGVTLIALANSMRTMRI
```

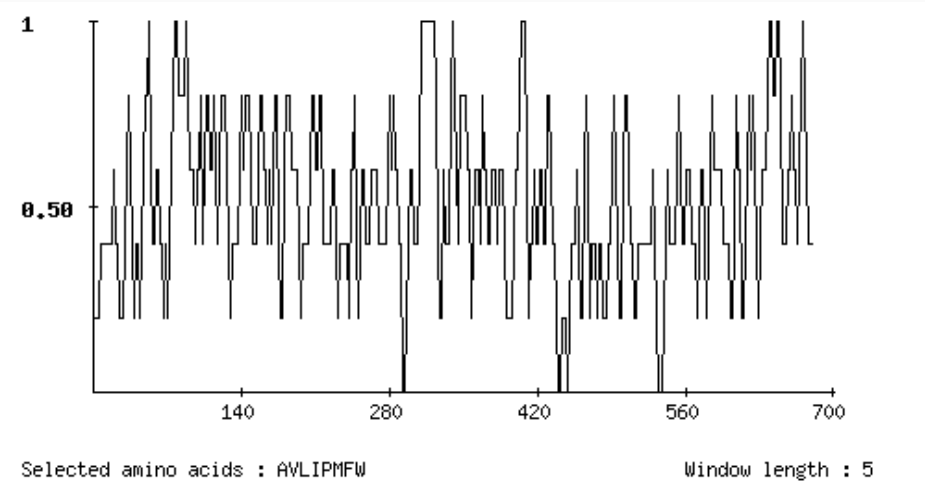
Helicobacter felis CadA

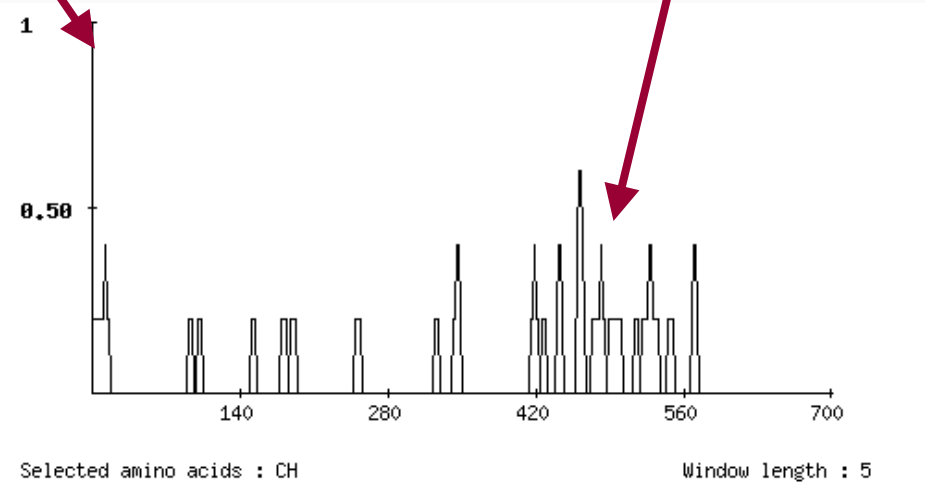
Cytoplasmic Domain responsible for Metal Ion Binding

N-terminal Domain with CH Cluster

Amino acides sélectionnés : AVLIPMFW
Largeur de la fenêtre : 5
Longueur de la séquence : 681 amino acides

Amino acides sélectionnés : CH
Largeur de la fenêtre : 5
Longueur de la séquence : 681 amino acides

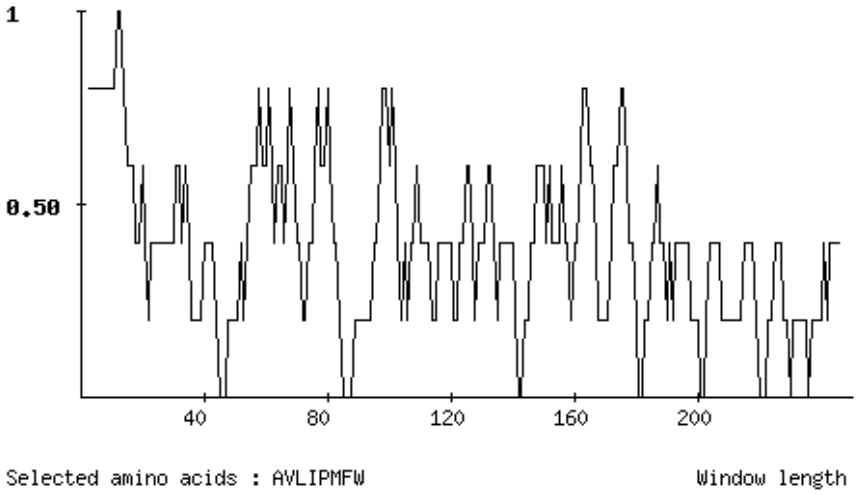




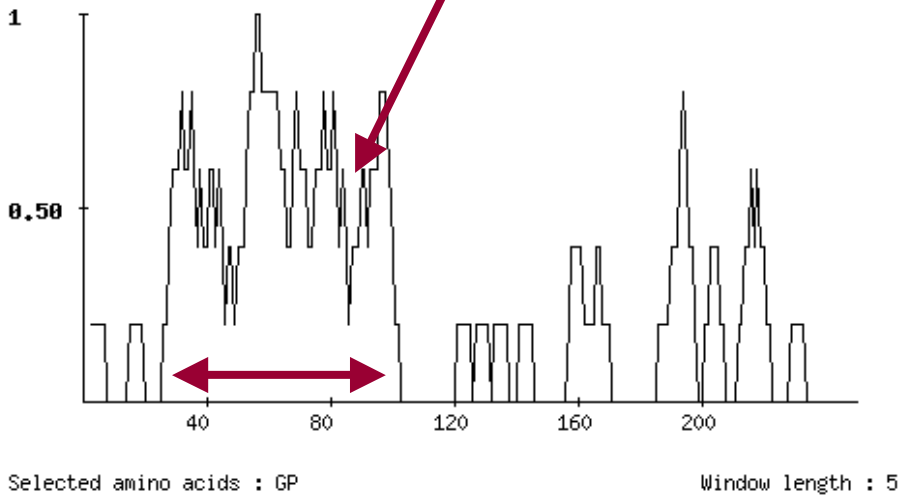
Human SP-A

The G and P-rich Collagenous Domain

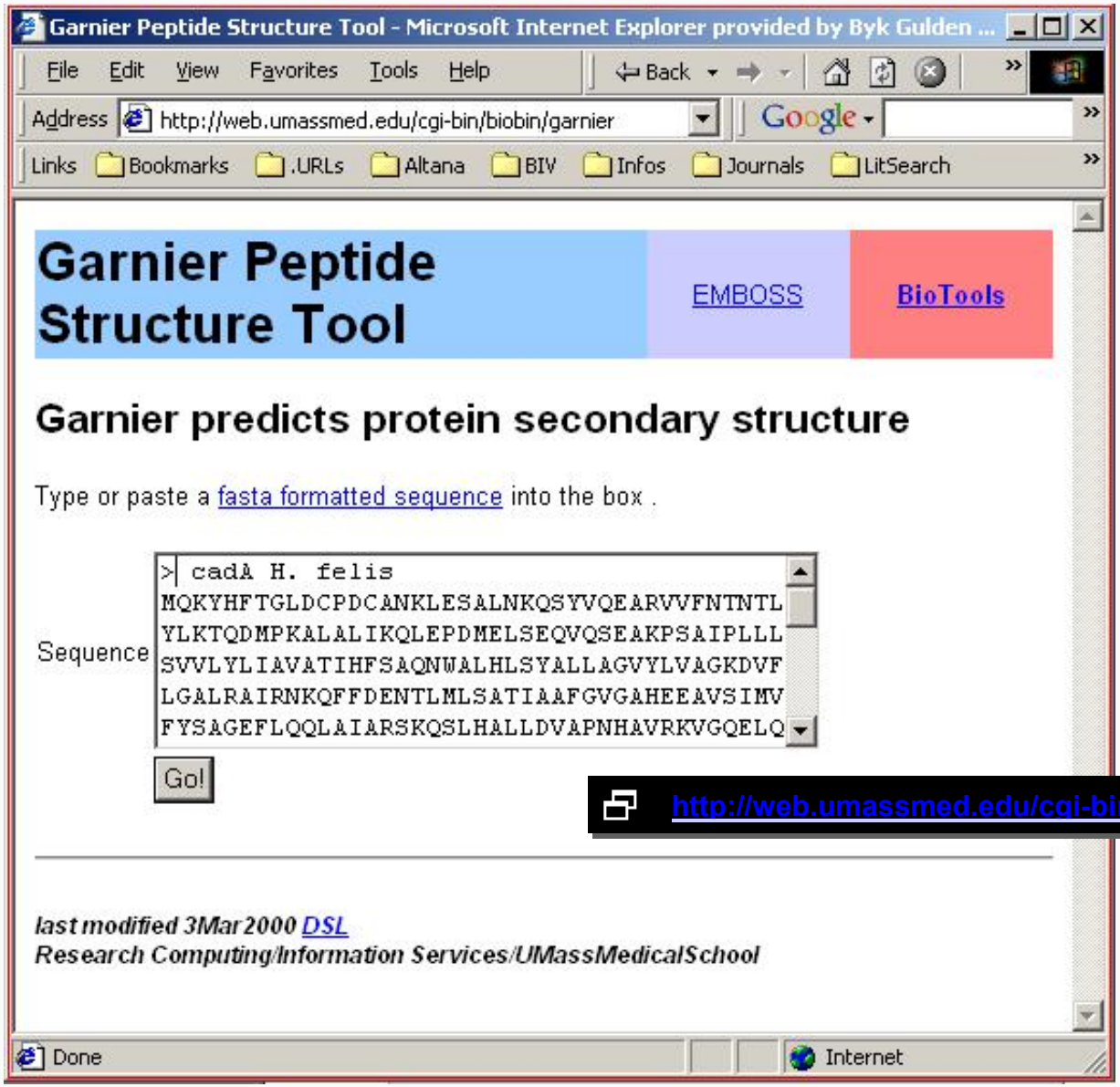
Amino acides sélectionnés : AVLIPMFW
Largeur de la fenêtre : 5
Longueur de la séquence : 248 amino acides



Amino acides sélectionnés : GP
Largeur de la fenêtre : 5
Longueur de la séquence : 248 amino acides



Looking for Secondary Structure Elements in CadA



http://web.umassmed.edu/cgi-bin/biobin/garnier



This Program presents its Results in Text Forrnat

```
GARNIER plot of CadA, 681 aa; DCH = 0, DCS = 0

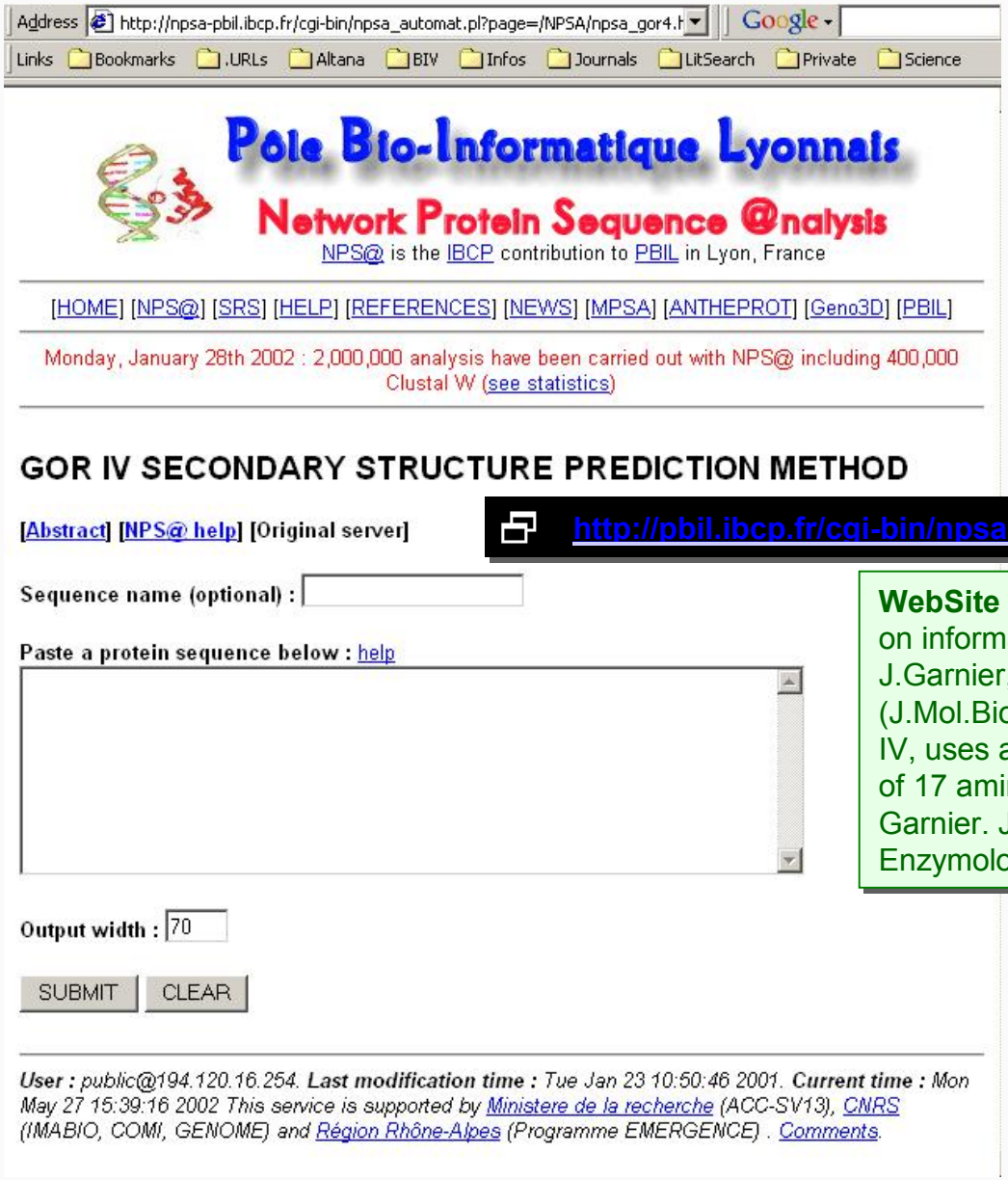
  Please cite:
  Garnier, Osguthorpe and Robson (1978) J. Mol. Biol. 120:97-120

           .   10    .   20    .   30    .   40    .   50    .   60
       MQKYHFTGLDCPDCANKLESALNKQSYVQEARVVFNTNTLYLKTQDMPKALALIKQLEPD
  helix HH           HHHHHHHHHHH                    HHHHHHHHHHHHHHH
  sheet   E                      EEEEEEEEE      EEEE
  turns    TTTTTTT TT           T           T   T
  coil            C              C         C CCC

           .   70    .   80    .   90    .  100    .  110    .  120
       MELSEQVQSEAKPSAIPLLLSVVLYLIAVATIHFSAQNWALHLSYALLAGVYLVAGKDVF
  helix HHHHHHHHHH      H         HHHHHH        HHHH   HHHH       HHHHH
  sheet               EE EEEEEEEE      EE          EEE    EEEEEE
  turns             T                       T
  coil            CCC                   CCCC
```

Note

A Graphically Formatted Program Server



http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html

WebSite Infos - GOR: The GOR method is based on information theory and was developed by J.Garnier, D.Osguthorpe and B.Robson (J.Mol.Biol.120,97, 1978). The present version, GOR IV, uses all possible pair frequencies within a window of 17 amino acid residues and is reported by J. Garnier. J.F. Gibrat and B.Robson in Methods in Enzymology, vol 266, p 540-553 (1996).

Secondary Structure Elements by Garnier's Method

GOR4 result for : UNK_18722080

Abstract GOR secondary structure prediction method version IV, J. Garnier, J.-F. Gibrat, B. Robson, Methods in Enzymology,R.F. Doolittle Ed., vol 266, 540-553, (1996)

View GOR4 in: [MPSA (Mac, UNIX) , About...] [AnTheProt (PC) , Download...] [HELP]

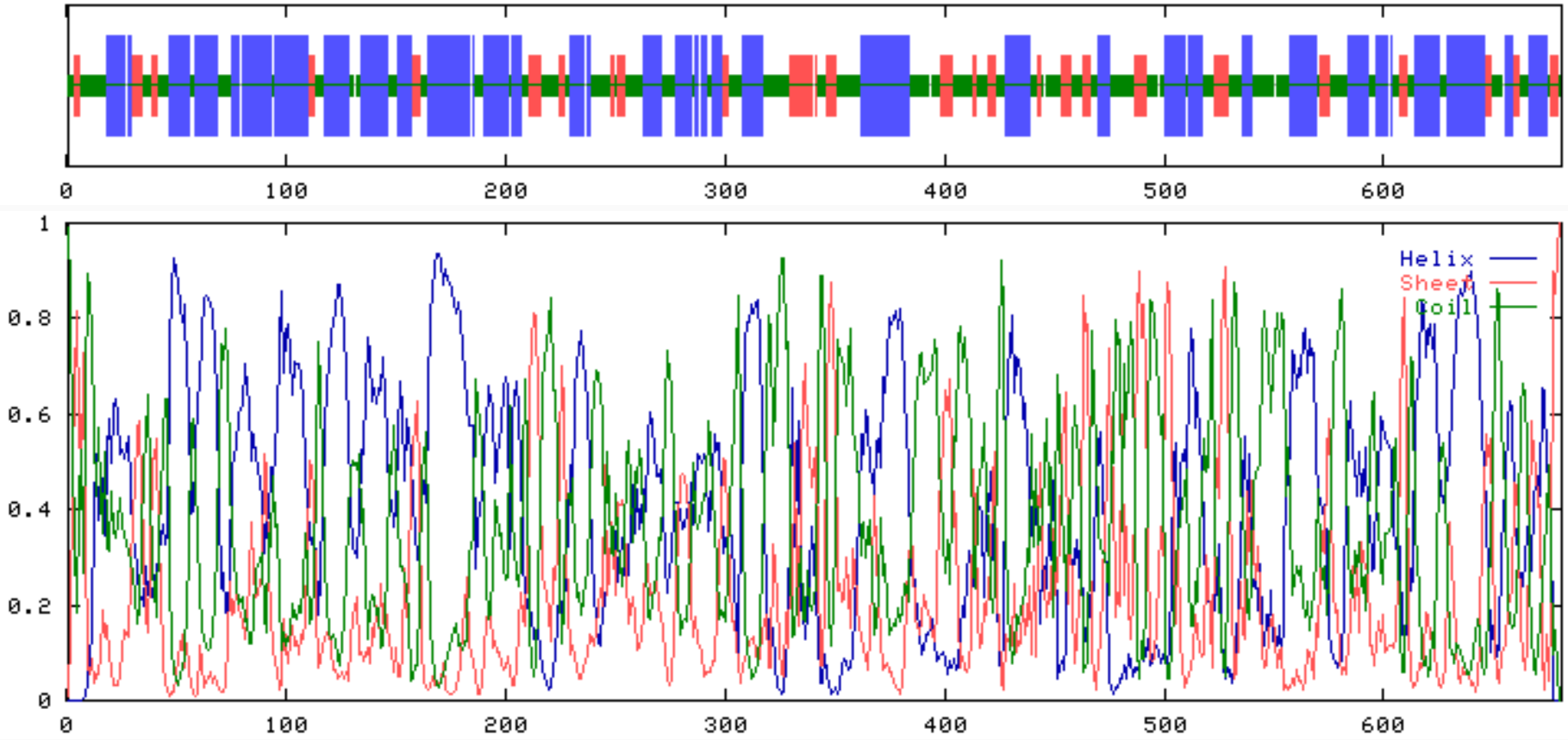
```
         10         20         30         40         50         60         70
          |          |          |          |          |          |          |
MQKYHFTGLDCPDCANKLESALNKQSYVQEARVVFNTNTLYLKTQDMPKALALIKQLEPDMELSEQVQSE
ccccceeeccccccccccchhhhhhhhhhheeeeecccceeecccccchhhhhhhhhhcchhhhhhhhhhh
AKPSAIPLLLSVVLYLIAVATIHFSAQNWALHLSYALLAGVYLVAGKDVFLGALRAIRNKQFFDENTLML
ccccccchhhhhhhhhhhhhhhhhhchhhhhhhhhhhhhhhheeecccchhhhhhhhhhhhcccchhhhhh
SATIAAFGVGAHEEAVSIMVFYSAGEFLQQLAIARSKQSLHALLDVAPNHAVRKVGQELQEVEPSALAIG
hhhhhhhcccchhhhhhheeeecccchhhhhhhhhhhhhhhhhhhhhcccchhhhhhhhhhhhchhhhhcc
DVVIVKVGEKIPTDGVVLRGESLLDQKALTGESLPVNVKEHSAVLGGSINLKGVLEIQVSKLYADSSVAK
ceeeeeecccccccceeeccchhhhhhhhhccccccccceeceeeecccccccchhhhhhhhhccccccchh
IVDLVSNAVSQKSQTEKFITTFARYYTPAVFAIALLIALVPPLLGHGDFDTWIYRGLFALMVSCPCALVI
hhhhhhchhhhhccchhhhheeecccccchhhhhhhhhhcccccccccccceeeeeeeeeeeeccccceeee
SVPLGYFGGVGAASRAGILIKSVQTLEALSQVKNIAFDKTGTLSKGEFNVIDVVPIAPFSKEDVLQHATC
ecccccccccccchhhhhhhhhhhhhhhhhhhhhhhcccccccccccccceeeeeecccccccccceecccc
AQILSTHPIAISIQKAYKRQCQHEVKDYQEIGGLGVQASCHSHLIIAGNDKMLHKYNIPHDTCSLEGTIV
eeeecccchhhhhhhhhhhhcccceeccccccceeeeecccccceeeeccchhhhhhcccccccccccceee
HVAVDGKHIGYIVVADTLKDNAKECLDGLKHAGIEHMCILSGDHEYSTKRVAKELDCPYYANLLPEDKLN
eeecccccccchhhhhhhhhhchhhhhhhcccccceeeeeeeccccccchhhhhcccccccccccccccchhh
AFKDFQAQHAHKSMFVGDGINDAPTLARADVSMSMGSASQISKESADIVITNNSLESVLKVFKIAKKTKR
hhhhhhhhhhceeeeecccccccchhhhhhhhhhcccchhhhhhhcccceeeeccchhhhhhhhhhhhcccch
IIIENIIFALAIKAMFIVLGLSGDASLWEAVLGDVGVTLIALANSMRTMRI
hhhhhhhhhhhhhhhhhheeeccccchhhheeeccccchhhhhhhhhceeeec
```

Sequence length : 681

```
GOR4 :
   Alpha helix      (Hh) :   318 is  46.70%
   3_10 helix       (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :   108 is  15.86%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :   255 is  37.44%
   Ambigous states  (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```

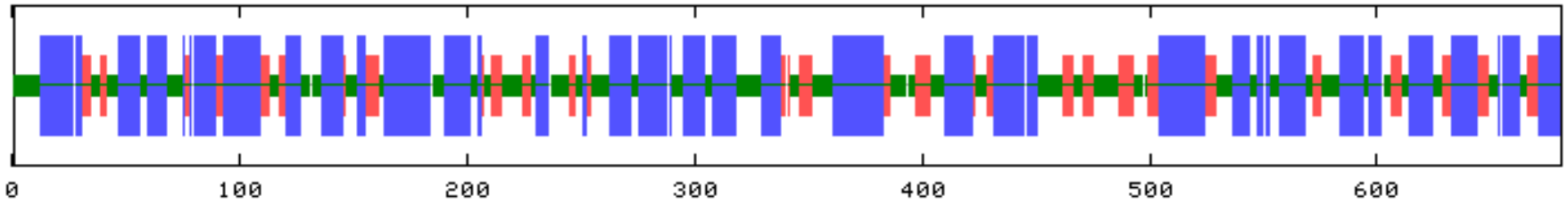
Note

Graphic Output is in BarGraph and PlotGraph Mode

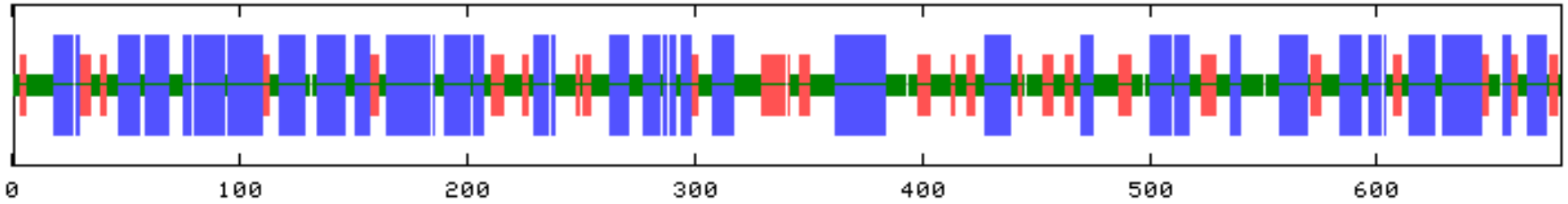


Similar Methods give Similar but not Identical Results

Geourjon



Garnier



Note

A BLOCKS Search Example

http://www.blocks.fhcrc.org/



WebSite Infos - Blocks WWW Server

A service for biological sequence analysis at the Fred Hutchinson Cancer Research Center in Seattle, Washington, USA. Visit the Blocks mirror site at the Weizmann Institute of Science in Israel.

```
BLIMPS (BLocks IMProved Searcher)   Version 3.2.5  1998/12
(C) Copyright 1993-8, Fred Hutchinson Cancer Research Center
Probe Sequence: ,  681 aa.
Probe Size: 681 Amino Acids
Target File (s) : /local/gensoft/lib/blimps/db/blocks.dat
Records Searched:   4034
Scores Done:        2863088
Alignments Done:    2863088

AC#           Description                                Strength  Score RF    AA#
BL00154F    0 E1-E2 ATPases phosphorylation site proteins.     1608   1422  0    572
              SmFVGDGINDAPtLARADVSMSMG
BL00154C    0 E1-E2 ATPases phosphorylation site proteins.     1466   1225  0    381
              VKnIAFDKTGTLsKGeFnV
BL00154A    0 E1-E2 ATPases phosphorylation site proteins.     1268   1219  0    209
              GDVVIVKVGEKIPTDGVV
BL01228D    0 Hypothetical cof family proteins.               1437   1184  0    573
              MfvGDGiNDAPtLarADVSmSMGSA
BL01047A    0 Heavy-metal-associated domain proteins.         1285   1161  0      5
              fTGLDCPDCANKLESALNKQSYV
BL00154B    0 E1-E2 ATPases phosphorylation site proteins.     1318   1131  0    343
              CPCALVISVPLGyfGGVGAASRAGILIKSvQtLEALS
BL00154G    0 E1-E2 ATPases phosphorylation site proteins.     1346   1111  0    597
              ASQISKESADIVITNNSLeSVLKVfKIAKKTkRI
BL01047B    0 Heavy-metal-associated domain proteins.          1118   1082  0    581
              DAPTLARADVSMSMGSASQIS
BL01049A    0 Uncharacterized protein family UPF0031 protei    1367   1060  0    518
              lKhagieHMCILsGDHEYstk
BL00555B    0 Plant viruses icosahedral capsid proteins 'S'    1363   1052  0    274
              DSSVAkIvDLVSNAVSQKsqteKFITtFARYy
```



DIGEST Gives You the Proteolytic Peptide Pattern of a Protein

```
DIGEST of CadA, from 1 to 681 Molwt= 73747.363

Complete digestion with Trypsin yields 47 fragments:
Start    End      Molwt       Sequence (up to 38 residues)
56       117      6721.838    (K) QLEPDMELSEQVQSEAKPSAIPLLLSVVLYLIAVATIH (D) ...
131      175      4895.566    (K) QFFDENTLMLSATIAAFGVGAHEEAVSIMVFYSAGEFL (S) ...
305      335      3507.125    (R) YYTPAVFAIALLIALVPPLLGHGDFDTWIYR (G)
644      677      3491.112    (K) AMFIVLGLSGDASLWEAVLGDVGVTLIALANSMR (T)
263      292      3161.634    (K) GVLEIQVSKLYADSSVAKIVDLVSNAVSQK (S)
336      365      2956.564    (R) GLFALMVSCPCALVISVPLGYFGGVGAASR (A)
447      471      2625.889    (K) DYQEIGGLGVQASCHSHLIIAGNDK (M)
412      435      2617.008    (K) EDVLQHATCAQILSTHPIAISIQK (A)
476      497      2368.640    (K) YNIPHDTCSLEGTIVHVAVDGK (H)
544      563      2356.669    (K) ELDCPYYANLLPEDKLNAFK (D)
4        24       2339.617    (K) YHFTGLDCPDCANKLESALNK (Q)
195      216      2293.636    (K) VGQELQEVEPSALAIGDVVIVK (V)
521      540      2284.544    (K) HAGIEHMCILSGDHEYSTKR (V)
178      194      1869.152    (K) QSLHALLDVAPNHAVRK (V)
628      643      1856.320    (K) TKRIIIENIIFALAIK (A)
604      620      1832.033    (K) ESADIVITNNSLESVLK (V)
396      411      1732.004    (K) GEFNVIDVVPIAPFSK (E)
573      588      1663.862    (K) SMFVGDGINDAPTLAR (A)
589      603      1498.683    (R) ADVSMSMGSASQISK (E)
293      304      1428.604    (K) SQTEKFITTFAR (Y)
217      229      1382.619    (K) VGEKIPTDGVVLR (G)
498      509      1328.570    (K) HIGYIVVADTLK (D)
250      262      1324.496    (K) EHSAVLGGSINLK (G)
33       43       1311.540    (R) VVFNTNTLYLK (T)
372      383      1302.487    (K) SVQTLEALSQVK (N)
 .        .          .         .   .
 .        .          .         .   .
 .        .          .         .   .
 .        .          .         .   .
```

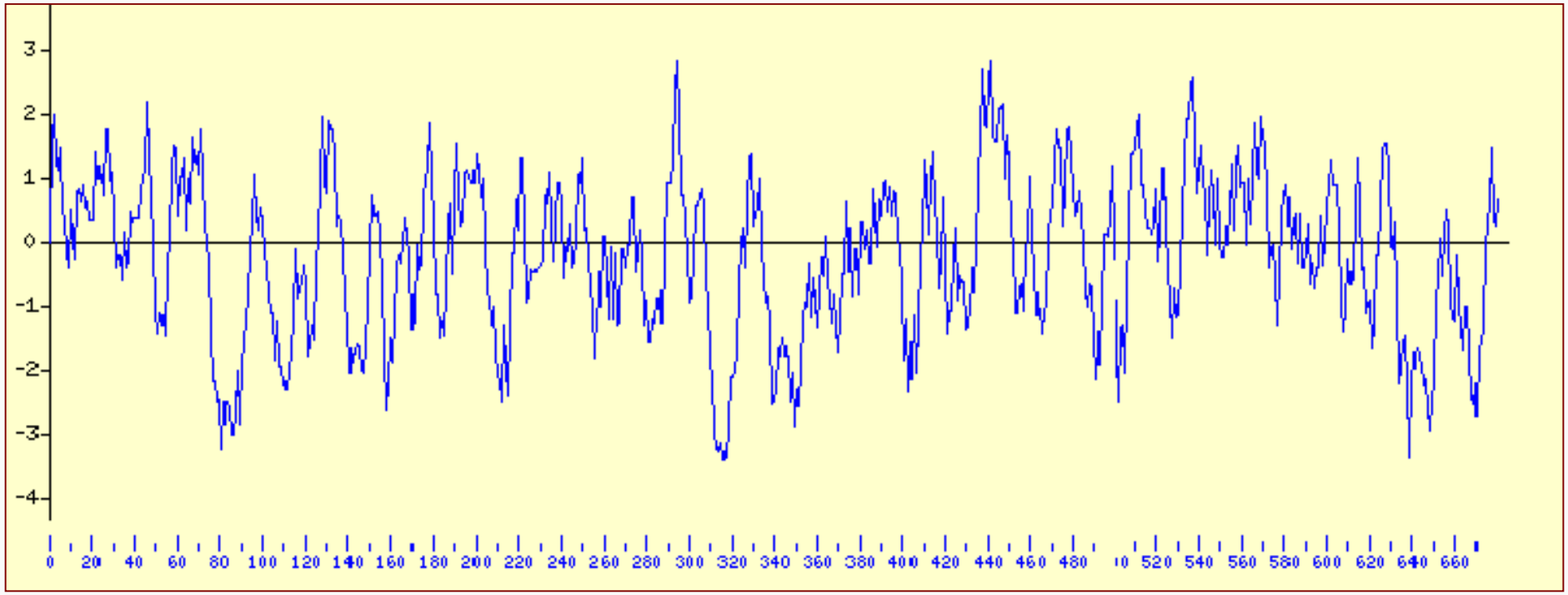
WebSite Infos - Program DIGEST

Protein proteolytic enzyme or reagent cleavage

DIGEST finds the positions where a specified proteolytic enzyme or reagent might cut a peptide sequence. This programs allows you to input a protein sequence and to specify one proteolytic agent from a list. It will then output a file containing the positions where the agent cuts, together with the peptides produced.

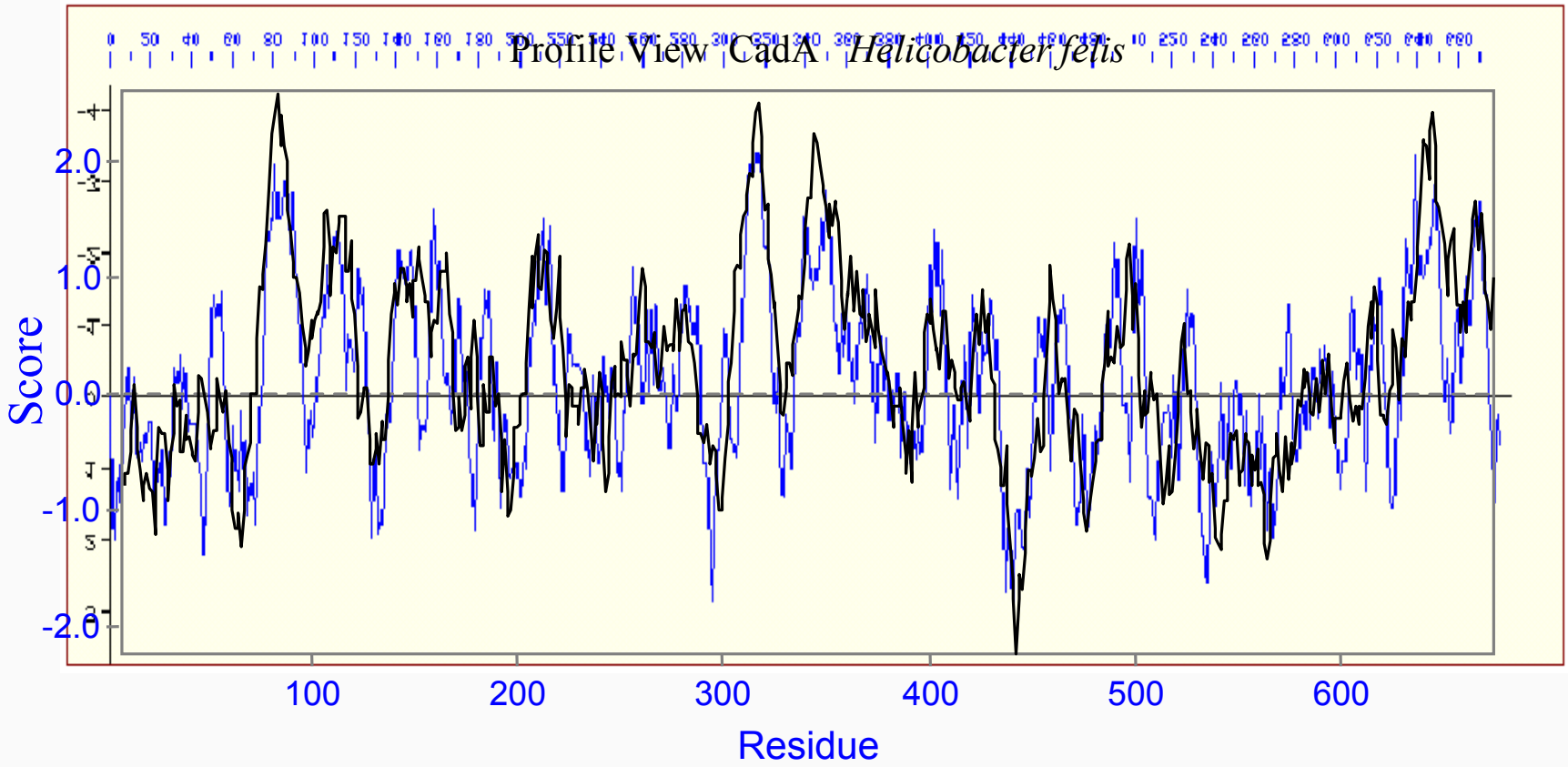
http://bioweb.pasteur.fr/seqanal/interfaces/digest.html

Hydropathy Profile of CadA According to Kyte & Doolittle



Note

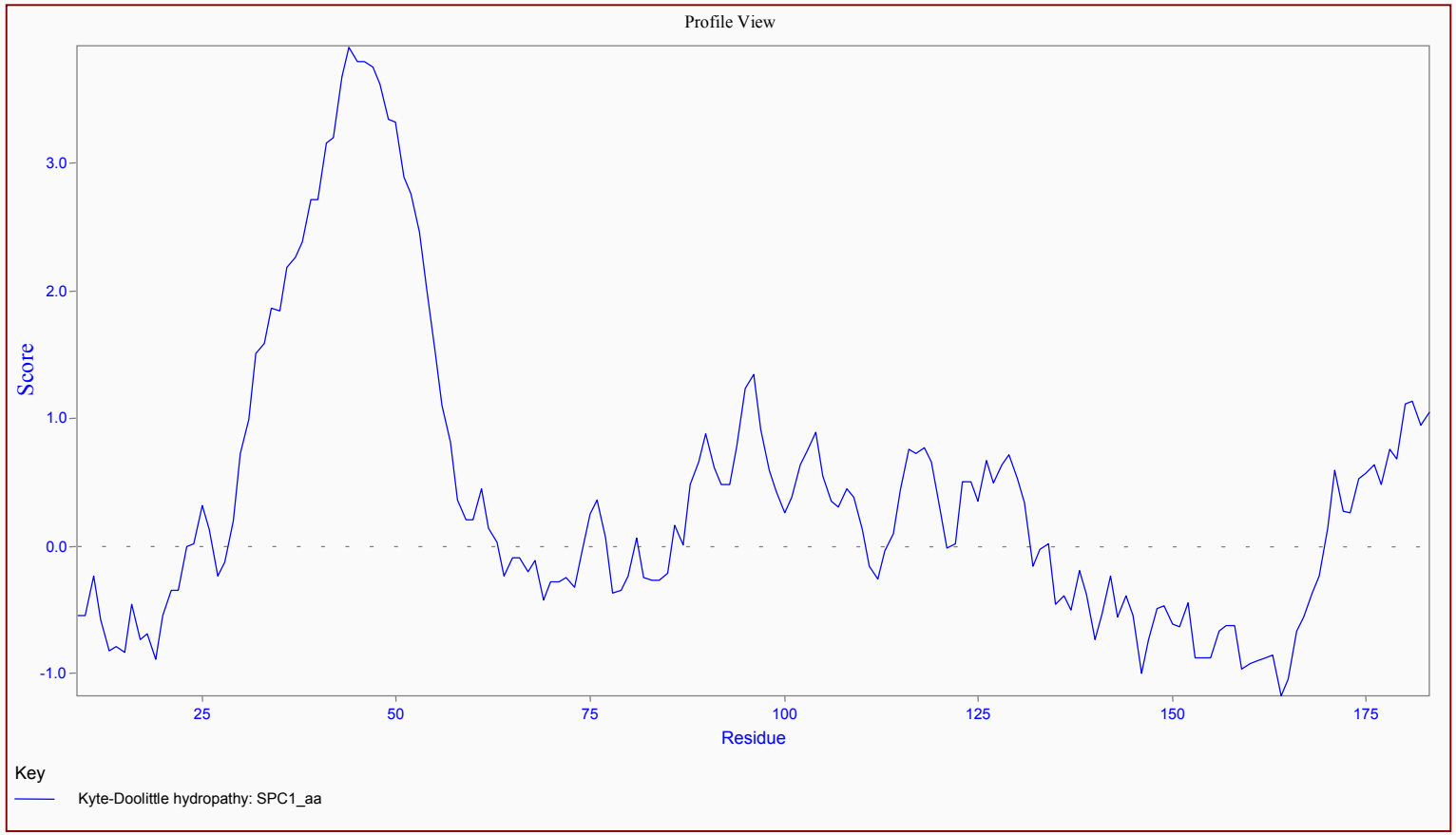
Hydrophobicity can be defined via Minima or Maxima



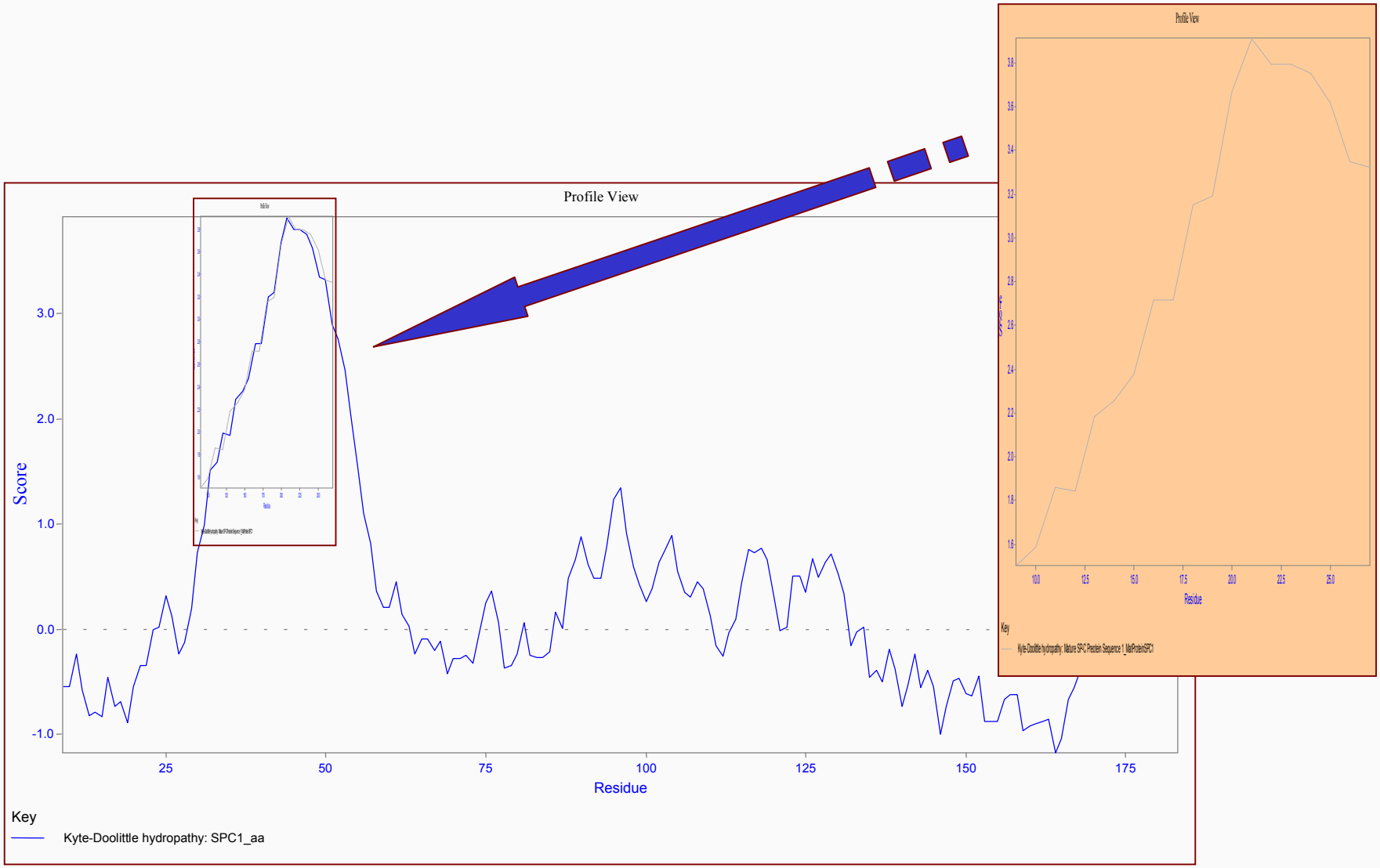
——— Kyte-Doolittle hydropathy: HfelisCadA_aa

Note

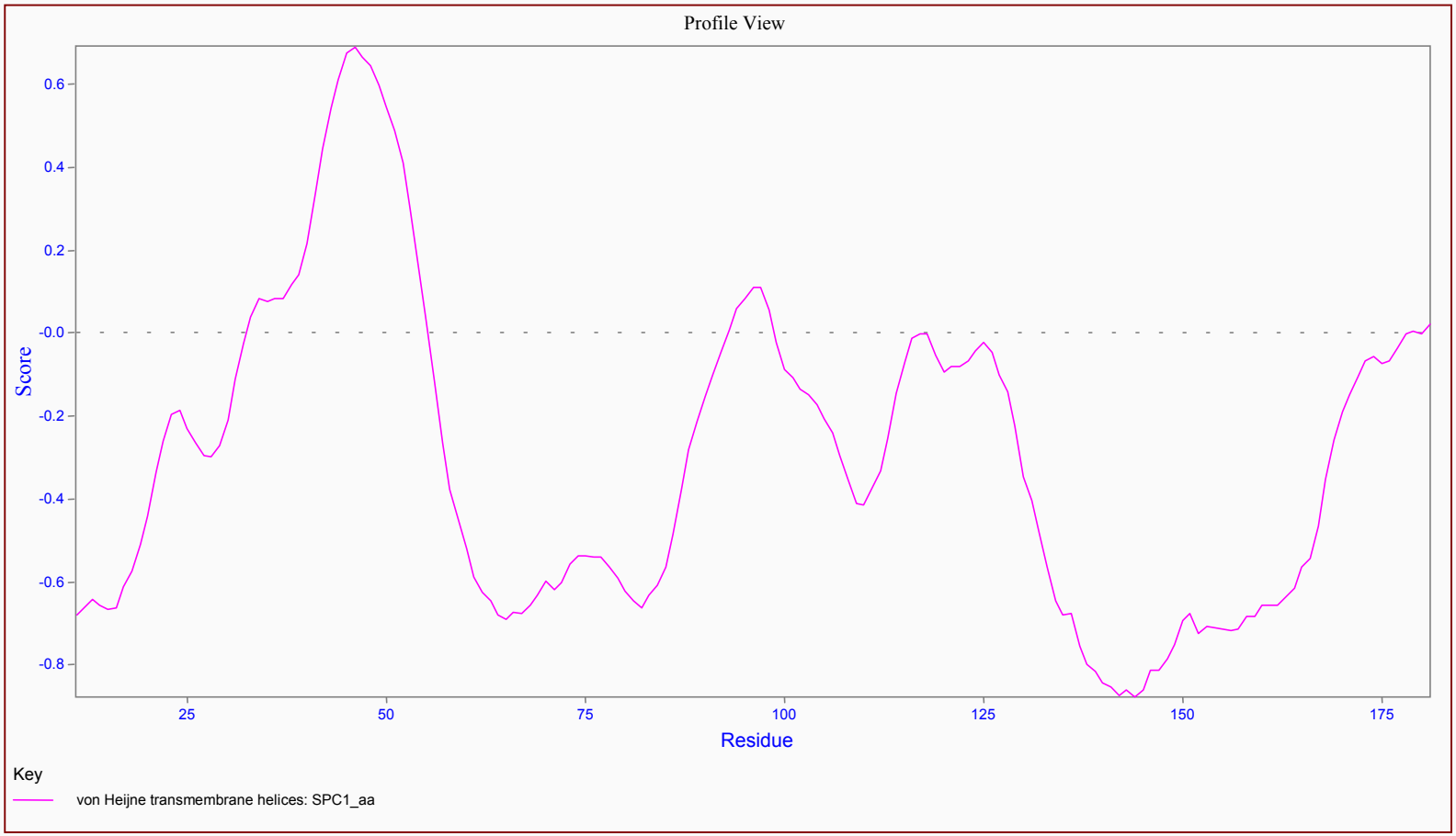
Hydropathy Profile of pre-SP-C (Kyte & Doolittle)



Hydropathy allows Localization of the mature SP-C Sequence

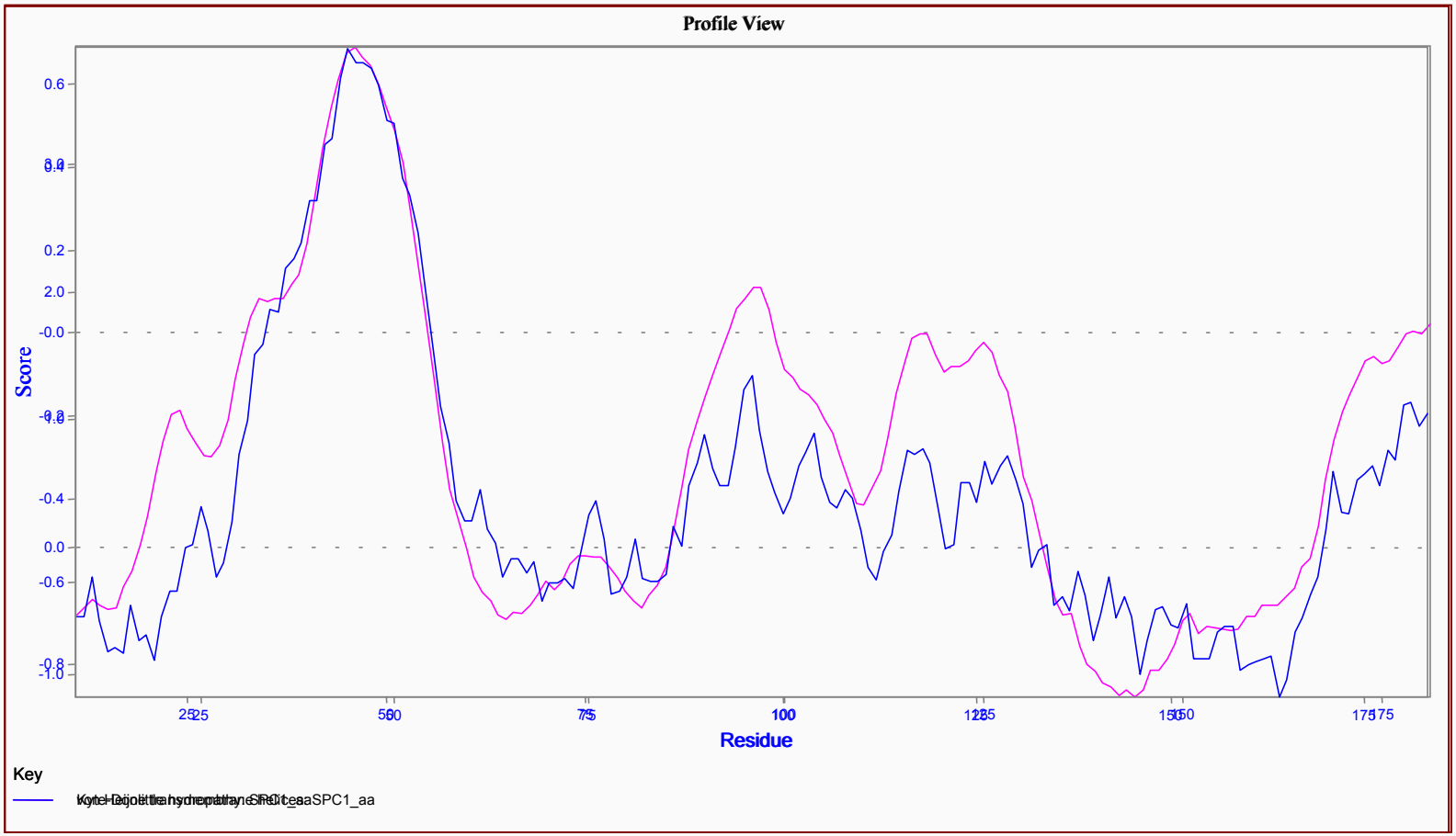


Antigenicity Profile of pre-SP-C according to Welling

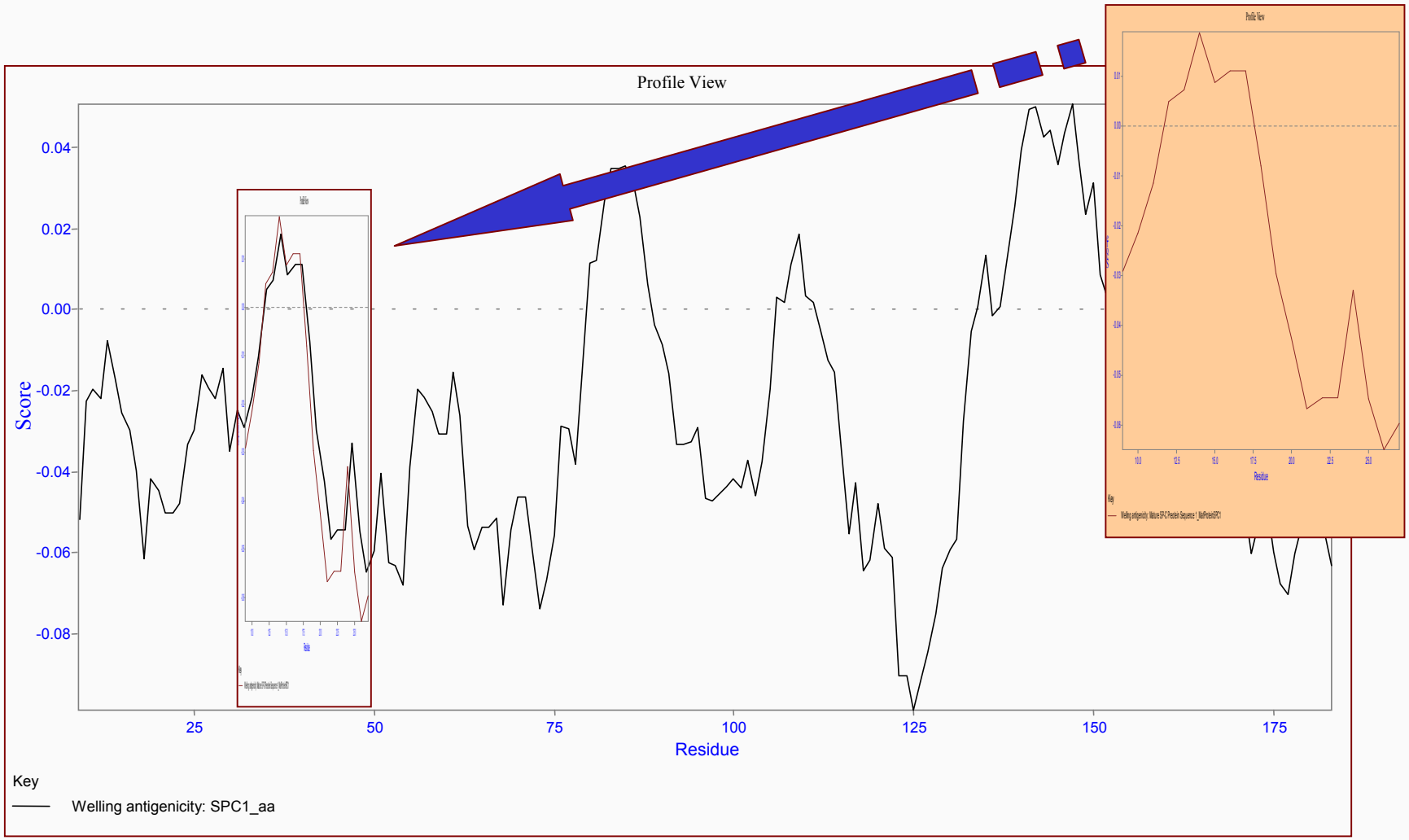


Note

Hydropathy and Antigenicity have similar Profiles



... aligning the SP-C Sequence with the Antigenicity Profile



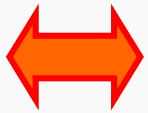


COG DataBase

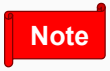
- What are COGs?
  - ➔ COG stands for Cluster of Orthologous Groups of proteins. The proteins that comprise each COG are assumed to have evolved from an ancestral protein, and are therefore either orthologs or paralogs.



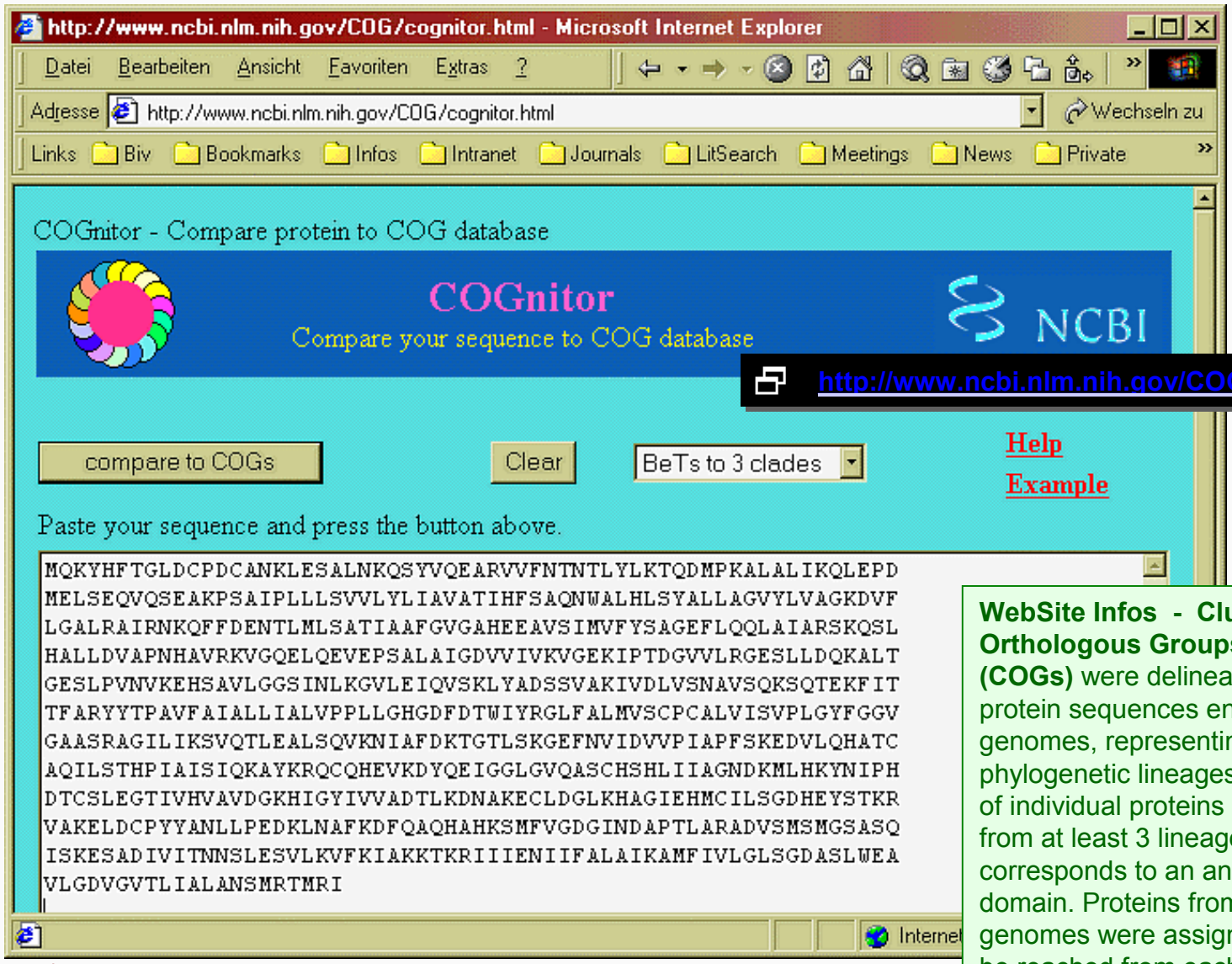
  - Orthologs are proteins from different species that evolved by vertical descent (speciation), and typically retain the same function as the original.



  - Paralogs are proteins from within a given species that are derived from gene duplication, and may evolve new functions that are related to the original.



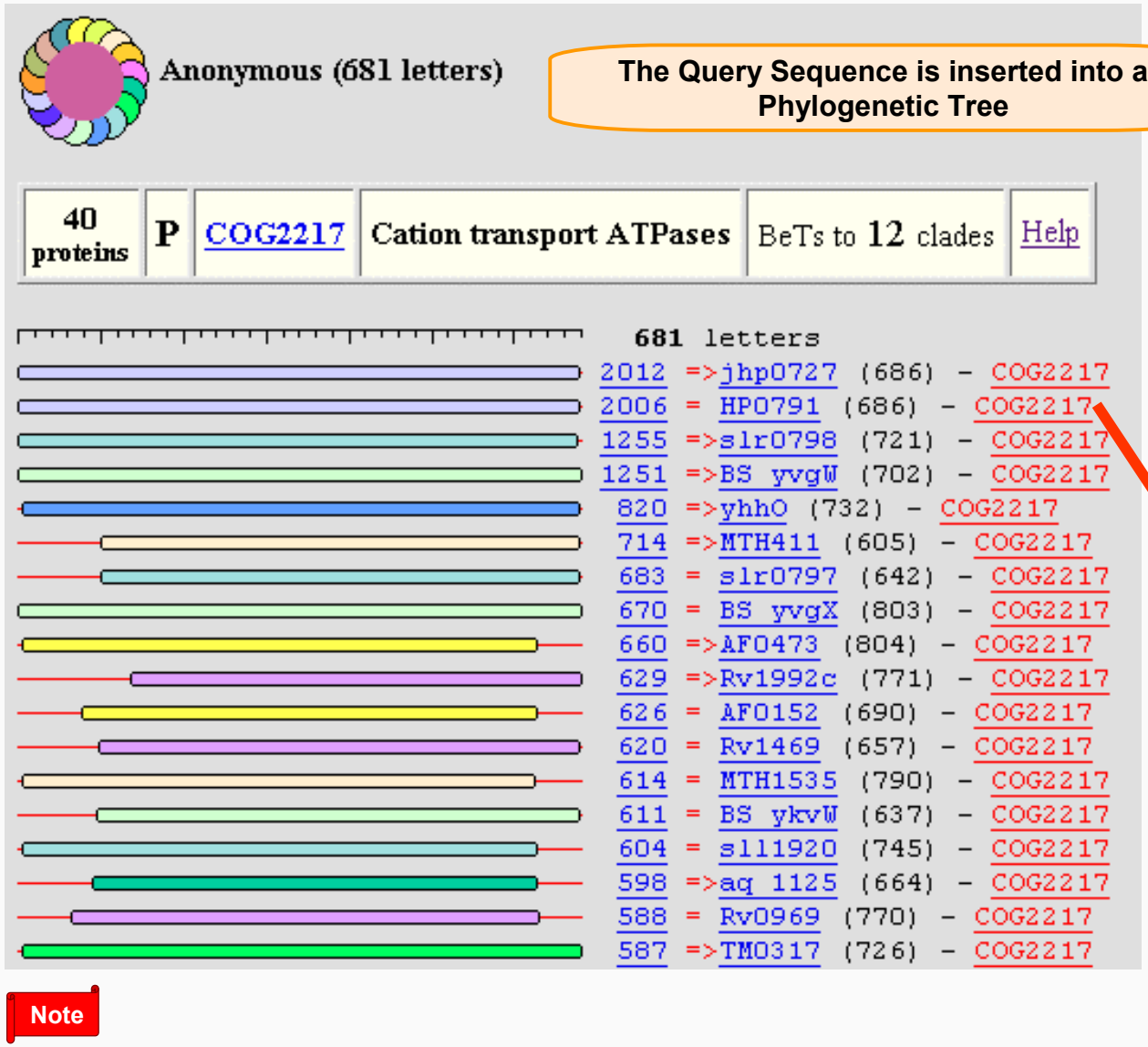
The COG Query Input Window (CadA)



WebSite Infos - Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 44 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain. Proteins from two eukaryotic genomes were assigned to COGs and can be reached from each individual COG page.

Note

Defining CadA - correctly - as a Cation Transport ATPase





From <Almost Identical> to <Barely Related>

```
# >HP1072
#           Length = 745
#
#  Score = 1352 bits (3462), Expect = 0.0
#  Identities = 694/745 (93%), Positives = 694/745 (93%)
#
# Query: 1   MKESFYIEGMTCTACSSGIERSLGRKSFVKKIEVSLLNKSANIEFDENQTNLDEIFKLIE 60
#            MKESFYIEGMTCTACSSGIERSLGRKSFVKKIEVSLLNKSANIEFDENQTNLDEIFKLIE
# Sbjct: 1   MKESFYIEGMTCTACSSGIERSLGRKSFVKKIEVSLLNKSANIEFDENQTNLDEIFKLIE 60
#
# Query: 61  KLGYSPKKALTKEKKEFFSPNVKLALAVIFTLFVVYLSMGAMXXXXXXXXXXXAIDNHSN 120
#            KLGYSPKKALTKEKKEFFSPNVKLALAVIFTLFVVYLSMGAM           AIDNHSN
# Sbjct: 61  KLGYSPKKALTKEKKEFFSPNVKLALAVIFTLFVVYLSMGAMLSPSLLPESLLAIDNHSN 120
#
# Query: 121 FLNACLQLIGALIVMHLGRDFYIQGFKALWHRQPNMSSLIAIGTSAALISSLWQLYLVYT 180
#            FLNACLQLIGALIVMHLGRDFYIQGFKALWHRQPNMSSLIAIGTSAALISSLWQLYLVYT
# Sbjct: 121 FLNACLQLIGALIVMHLGRDFYIQGFKALWHRQPNMSSLIAIGTSAALISSLWQLYLVYT 180
#
# Query: 181 NHYTDQW
#            NHYTDQW
# Sbjct: 181 NHYTDQW
```

```
# >MJ1377
#           Length = 553
#
#  Score = 29.7 bits (65), Expect = 9.5
#  Identities = 20/38 (52%), Positives = 24/38 (62%), Gaps = 5/38 (13%)
#
# Query: 48 NQTNLDEIFKLIEKLGYSPKKALTKEKKE--FFSPNVK 83
#           N+ +LDEI K +E+LG   KK   KEKKE     PNVK
# Sbjct: 55 NKLSLDEIKKKLEELGEDVKK---KEKKEKGLELPNVK 89
```

Note