Strojno učenje prediktivnih modela u računalnoj biologiji

Tomislav Šmuc & Fran Supek

Strojno učenje prediktivnih modela u računalnoj biologiji

- Grupa za računalnu biologiju i bioinformatiku
  - T. Šmuc & F. Supek (ZEL)
  - Suradnici na projektu:
  - M. Kralj (IRB - ZMM)
  - B. Balen (PMF – Biologija)
  - B. Kojić-Prodić (IRB-FK)

  - + Interna suradnja (Zavod; IRB)
    - N. Škunca, D. Gamberger
    - J. Repar (grupa Zahradka)
    - M. Bošnjak
    - Grupa za ihtiobiologiju

Strojno učenje prediktivnih modela u računalnoj biologiji

Međunarodna suradnja:

- IJS – KT Dept (Sašo Džeroski).
  - Inductive databases in genomics and proteomics (HR-SI bilaterala)

Neformalno

- ETH – Comp. Bioch. Res. Group (Cristophe Dessimoz)
- MedILS / INSERM (Anita Kriško)

Drugi projekti (vezani uz suradnike i temu projekta)

- Iprojekti
  - ReviGO (revigo.irb.hr)
  - GORBI (gorbi.irb.hr)

Strojno učenje prediktivnih modela u računalnoj biologiji

Teme projekta

- **(Bio-I) Bioinformatics *sensu stricto***
- **(Bio-II) Making experimental proteomics more reliable**
- **(Bio-III) Models for drug discovery**

- II godina projekta (startao 2007.)

  (Bio-I) Bioinformatics *sensu stricto*

  - Codon usage bias & translational selection in bacteria and archaea
    - (F. Supek - tema za doktorat)
  - Suradnja: Ortholog&paralogs in phylogenetic profiling
    (N. Škunca + IJS-KT grupa)

Strojno učenje prediktivnih modela u računalnoj biologiji

Codon usage bias & translational selection in bacteria and archaea – paper:

Translational selection is ubiquitous and reflects environmental adaptation in prokaryotes

Fran Supek, Nives Škunca, Jelena Repar, Kristian Vlahoviček, Tomislav Šmuc

Translational selection is ubiquitous .....

Translacija = proizvodnja proteina po uputi mRNA
(mRNA='fotokopija' gena)

1 kodon (3 'slova') = 1 aminokiselina

Neki kodoni su malo brži ili točniji u translaciji, da li to ostavlja trag u evoluciji bakterijskih genoma?

<u>Osnovne hipoteze</u>:

1. Pristup preko strojnog učenja je značajno bolji od pristupa baziranih na indeksima udaljenosti
2. OCU (optimirani odabir kodona)
    a) prisutan u svim genomima prokariota (a ***ne*** samo u brzorastućim)
    b) vezan za određene funkcije gena, odnosno prilagođen uvjetima okoliša
    c) jako koreliran uz ekspresiju gena (proxy for gene expression)

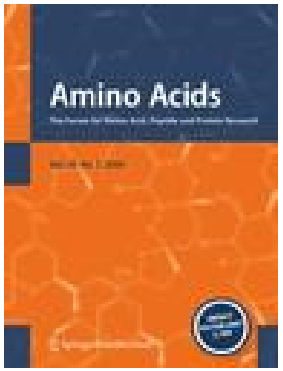
Models for drug discovery

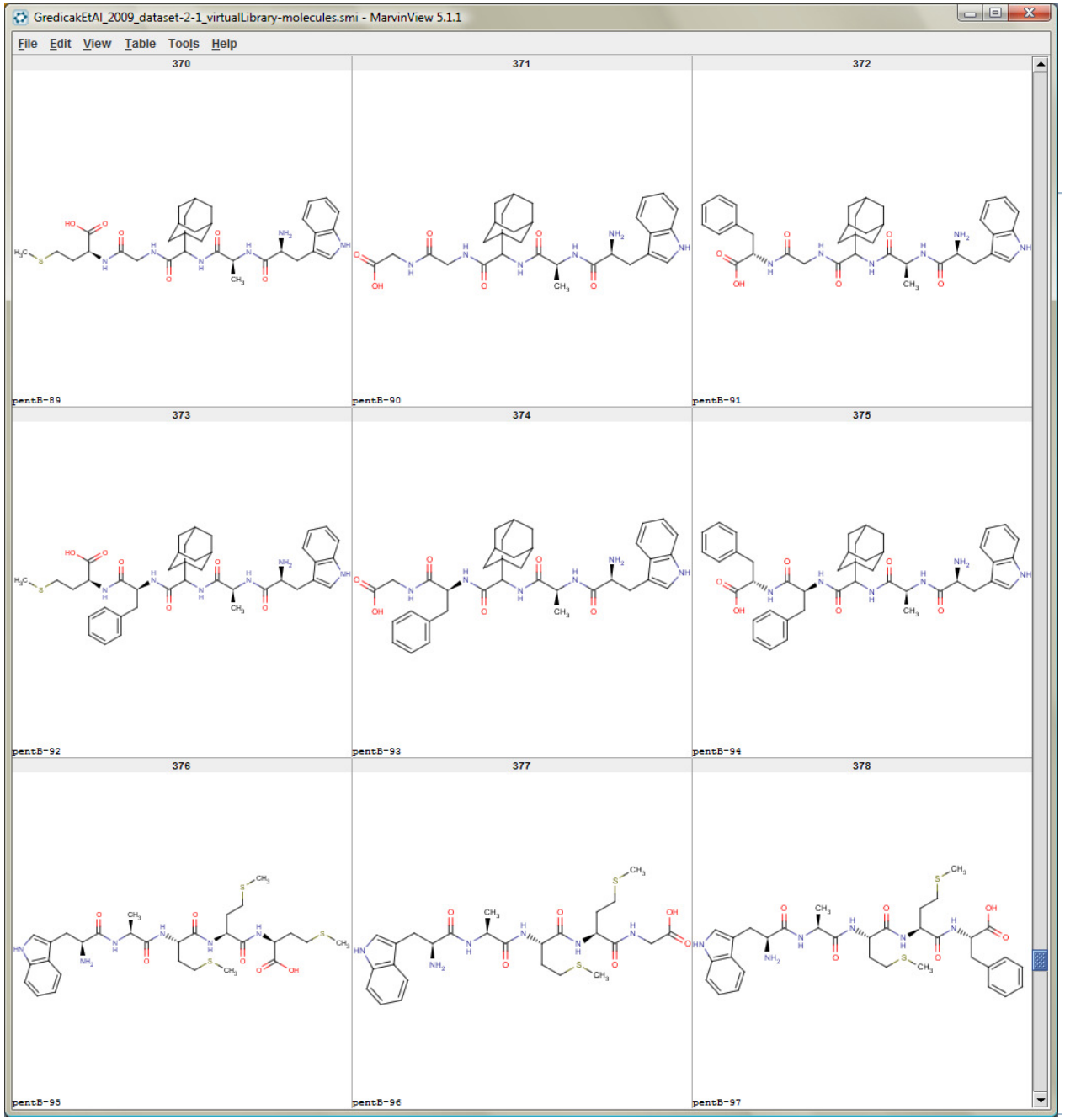
Computational structure-activity study directs synthesis of novel antitumor enkephalin analogs

M. Gredičak,[a] F. Supek,[b] M. Kralj,[c] Z. Majer,[d] M. Hollósi,[d] T. Šmuc,[b] K. Mlinarić-Majerski,[a] Š. Horvat[a]*

[a] Division of Organic Chemistry and Biochemistry, Ruđer Bošković Institute, Zagreb, 10002-HR Croatia
[b] Division of Electronics, Ruđer Bošković Institute, Zagreb, 10002-HR Croatia
[c] Division of Molecular Medicine, Ruđer Bošković Institute, Zagreb, 10002-HR Croatia
[d] Department of Organic Chemistry, Institute of Chemistry, Eötvös Loránd University, H-1518 Budapest, Hungary

I.F. = 4.1





Met-enkephalin:

Tyr-Gly-Gly-Phe-Met

22 peptides with known activity; virtual library = 390 adamanatane-containing peptides

Met-enkephalin:

Tyr-Gly-Gly-Phe-Met

1330 molecular descriptors (DRAGON software)

Support vector machines regression (non-linear)

22 training set molecules
q = 0.734 (usable)

Result: very hydrophobic peptids with good antitumor activity; problems with solubility.

Table 1. Top-rated peptides as obtained by computational cytostatic activity prediction, chosen from a virtual library of 390 enkephalin-like peptides containing the (*R*,*S*)-(1-adamantyl)glycine (Aaa) residue

Peptide	Predicted activity[a]	log*P*[b]	log*S*[b]
Tyr-Aaa-Gly-Phe-Met (**2**)[c]	-5.77	0.65	-5.36
Phe-Aaa-Gly-Phe-Met (**5**)	-5.37	0.91	-5.62
Phe-Aaa-Gly-Phe-Phe (**4**)	-5.07	1.32	-5.83
Tyr-Aaa-Gly-Phe-Phe (**3**)	-4.20	1.10	-5.65
Trp-Aaa-Gly-Phe-Met	-4.18	1.14	-5.79
Tyr-Aaa-Gly-Phe-Gly	-3.53	0.40	-5.02
Trp-Aaa-Gly-Phe-Phe	-3.50	1.58	-5.90
Phe-Aaa-Gly	-3.28	0.29	-4.68
Phe-Gly-Gly-Phe-Phe	-3.27	0.01	-5.23
Phe-Gly-Aaa-Gly	-3.25	-0.12	-4.47

[a] The predicted activity is dimensionless; more negative values signify stronger cytostatic activity.
[b] log*P* (hydrophobicity) and log*S* (aqueous solubility) values are given as predicted by the E-Dragon web service (Tetko et al., 2005).
[c] Horvat et al., 2006.

- Hvala na pažnji!