

ClustScan and *CompGen* Program Packages

Novalis *and*

Antonio Starčević





ClustScan program package:

Screenshot of the ClustScan software interface showing the workspace and annotation editor, along with a floating Import DNA wizard and HMMER parameters dialog.

Import DNA wizard (Top Left):

- Import sequences:**
 - Choose method:**
 - Import DNA
 - Import DNA
 - Import DNA
 - Import from clipboard
 - Search for genes with:**
 - Search for genes with
 - Search for genes with
 - Search for domains with HMMER:**
 - Stringent (E-value <= 10⁻³)
 - Relaxed (E-value < 10)

HMMER parameters (Bottom Right):

- Stringent (E-value <= 10⁻³)
- Relaxed (E-value < 10)

Workspace (Main Area):

Annotation editor: PKS 1

DNA sequence (PKS 1):

S R S I P R N * Y D S L * G N C E R I T I P S R R R
L D R S R E I N T T H Y R G I V S G * Q F P R E G D
* I D P A K L I R L T I G E L * A D N N S L E K E I
0 10 20 30 40 50 60 70
TCTAGATCGATCCCGCGAAATTAAATACGACTCACTATAGGGATTGTGAGCGGATAACAATTCCCTCGAGAAGGGAGATA
31531 31521 31511 31501 31491 31481 31471 31461
AGATCTAGCTAGGGCGCTTAATTATGCTGAGTGATATCCCCTAACACTCGCCTATTGTTAAGGGAGCTTCCCTCAT
R S R D R S I L V V * * L P I T L P Y C N G R S P S
* I S G A F N I R S V I P S N H A S L L E R S F S I
L D I G R F * Y S E S Y

File Menu:

- File
- Edit
- Search
- Help

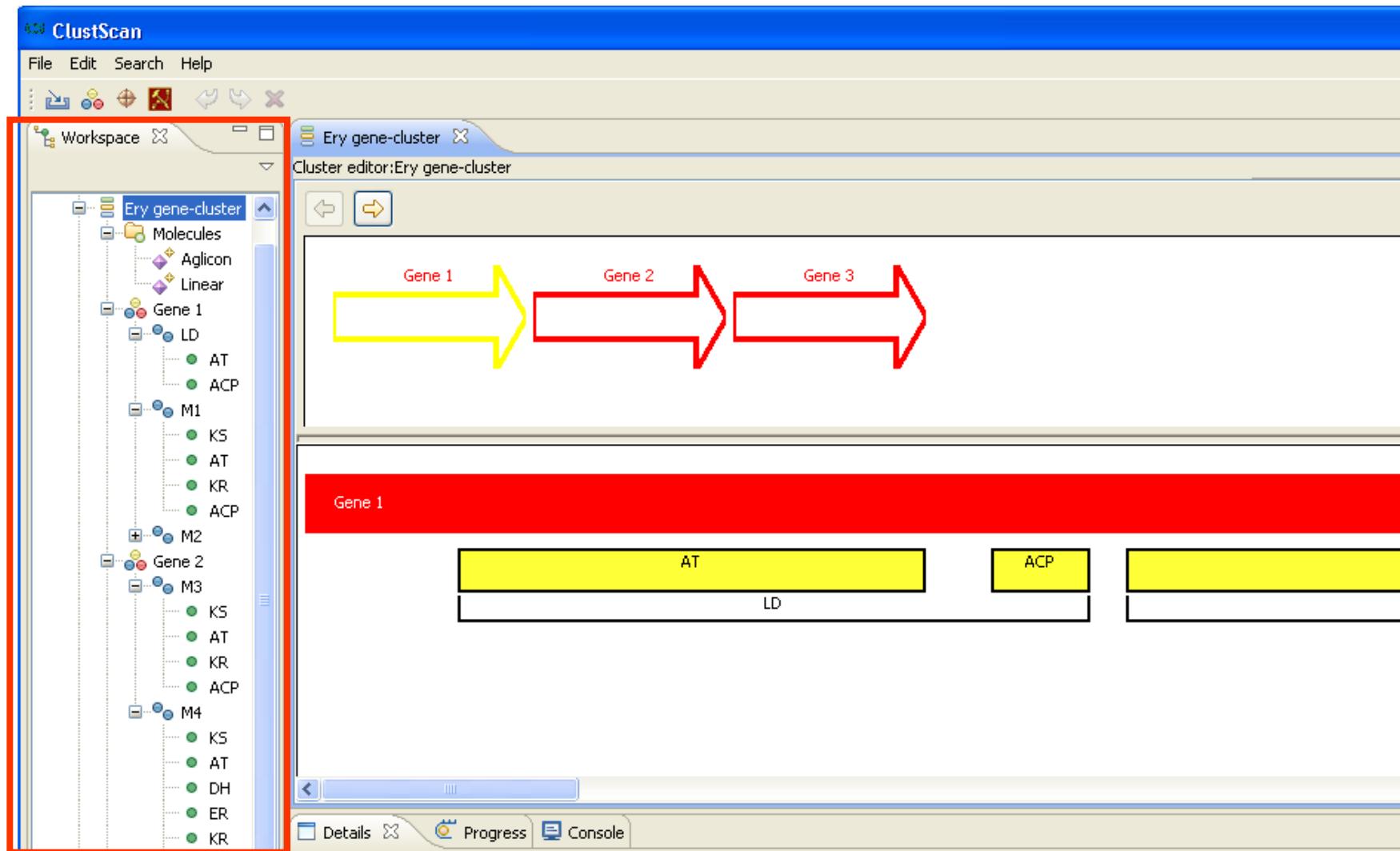
Toolbar:

- New
- Open
- Save
- Print
- Exit
- Find
- Replace
- Copy
- Paste
- Cut
- Delete
- Search
- Help

Bottom Bar:

- Setting ...
- DNA seq...
- Kalendar...
- 4th CEF...
- Microsoft...
- ClustScan
- ClustScan
- 18:08

Cluster editor and Workspace window:



Annotation editor:

ClustScan

File Edit Search Help

Workspace Ery gene-cluster Annotation editor: Ery gene-cl

SMILES:
CC[C@H](O)[C@H](C)[C@H](O)[C@H](C)C(=O)C(C)
C[C@H](C)[C@H](O)[C@H](C)[C@H](O)C(C)C(S)=O

The screenshot shows the ClustScan interface with the following details:

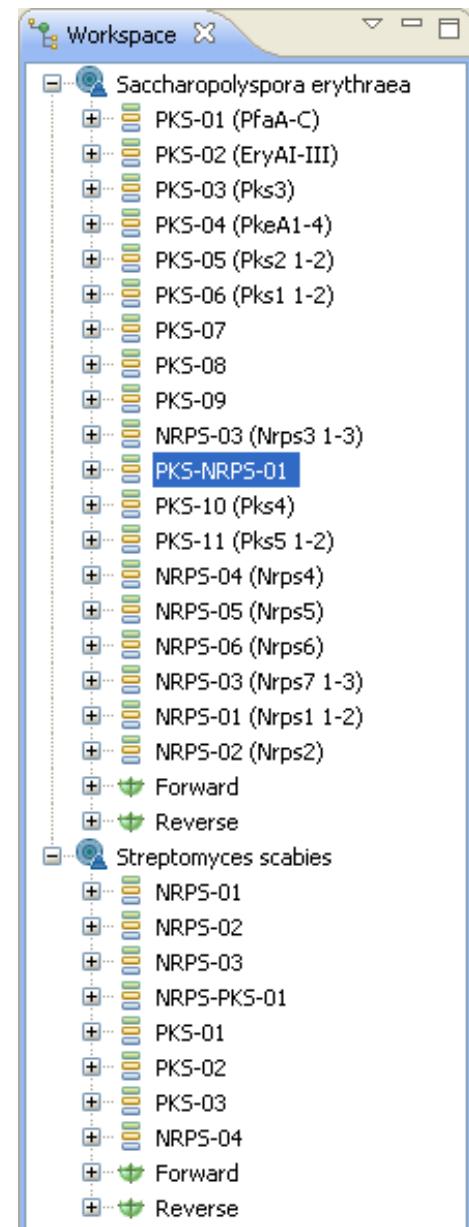
- Workspace:** Shows a tree view of the project structure under "Ery gene-cluster". The "Linear" node is selected and highlighted with a red box.
- Annotation editor:** Displays the SMILES string for the selected domain KR, which corresponds to the chemical structure shown below.
- Chemical Structure:** A 3D ball-and-stick model of the KR domain, showing a repeating motif of a beta-turn (CH₃-OH) connected by a carbonyl group (C=O). The structure is oriented vertically, with the first residue on the left and the last on the right.
- Domain Properties:** A detailed panel for the KR domain, also highlighted with a red box. It includes:
 - DNA coordinates: 14346..14808 (462 pb)
 - Protein frame: Forward 3
 - Protein coordinates: 4780..4934 (154 aa)
 - Score: 153.492
 - E-value: 4.35914E-46
 - Activity: Inactive
 - Specificity: Chirality of Me: S

Sequencing of *Sac. erythraea* genome:

2007 - Peter Leadley's group published DNA sequence of the 8.2 Mb *Sac. erythraea* genome (11 PKS, 6 NRPS, 1 PKS-NRPS)



(Oliynyk *et al*, *Nat. Biotechnol.*, 25, 447, 2007)



ClustScan results using *Sac. erythrea* genome:

Annotation editor: Sac. erythrea PKS gene-cluster

The screenshot displays the annotation editor interface for the *Sac. erythrea* PKS gene-cluster. The top navigation bar shows two tabs: "Sac. erythrea PKS gene-cluster" and "Annotation editor: Sac. erythrea PKS gene-cluster". Below the tabs, several windows are open, each showing domain properties for different domains:

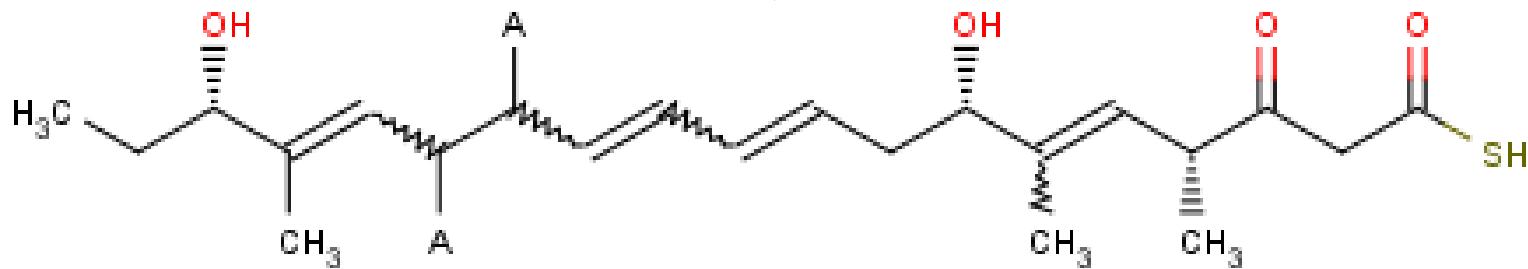
- ACP (Acyl carrier protein):** DNA coordinates: 7128..7329 (201 pb), Protein frame: Reverse 2, Protein coordinates: 14089..14156 (67 aa), Score: 68.41, E-value: 1.785E-20, Specificity: Chirality of Me: R, Chirality of OH: S.
- KR (Ketoreductase):** DNA coordinates: 7698..8184 (486 pb), Protein frame: Reverse 2, Protein coordinates: 13804..13966 (162 aa), Score: 231.299, E-value: 1.64878E-69, Activity: active, Specificity: Chirality of Me: R, Chirality of OH: S.
- AT (Acyl transferase):** DNA coordinates: 9108..10056 (948 pb), Protein frame: Reverse 2, Protein coordinates: 13180..13496 (316 aa), Score: 497.782, E-value: 9.94904E-150, Activity: active, Specificity: Prediction: methyl-malonyl.
- KS (Ketosynthase):** DNA coordinates: 10347..11622 (1275 pb), Protein frame: Reverse 2, Protein coordinates: 12658..13083 (425 aa), Score: 1127.0, E-value: 0.0, Specificity: Prediction: methyl-malonyl.
- TE (Thioesterase):** DNA coordinates: 1488..2097 (609 pb), Protein frame: Reverse 3, Protein coordinates: 15832..16035 (203 aa), Score: 115.659, E-value: 1.06726E-34, Specificity: Chirality of Me: R, Chirality of OH: S.
- ACP (Acyl carrier protein):** DNA coordinates: 43452..43653 (201 pb), Protein frame: Reverse 2, Protein coordinates: 1981..2048 (67 aa), Score: 63.287, E-value: 6.22033E-19, Specificity: Chirality of Me: R, Chirality of OH: S.
- AT (Acyl transferase):** DNA coordinates: 43842..44814 (972 pb), Protein frame: Reverse 2, Protein coordinates: 1594..1918 (324 aa), Score: 213.631, E-value: 3.43368E-64, Activity: active, Specificity: Prediction: Acetyl-CoA.
- KS (Ketosynthase):** DNA coordinates: 45120..46380 (1260 pb), Protein frame: Reverse 2, Protein coordinates: 1072..1492 (420 aa), Score: 750.12, E-value: 1.08763E-225, Activity: active, Specificity: Chirality of Me: R, Chirality of OH: S.

Below the domain properties, a large protein structure diagram is shown. The protein is composed of several domains, each represented by a colored box (green, yellow, red, blue) with a black outline. The domains are arranged in a linear fashion, with some domains having internal sub-domains indicated by smaller boxes. Arrows point from the domain names in the list above to their corresponding domains in the structure diagram. The structure diagram also features red and blue vertical bars and a wavy red line at the bottom.

ClustScan results using Sac. erythrea genome:

SMILES:

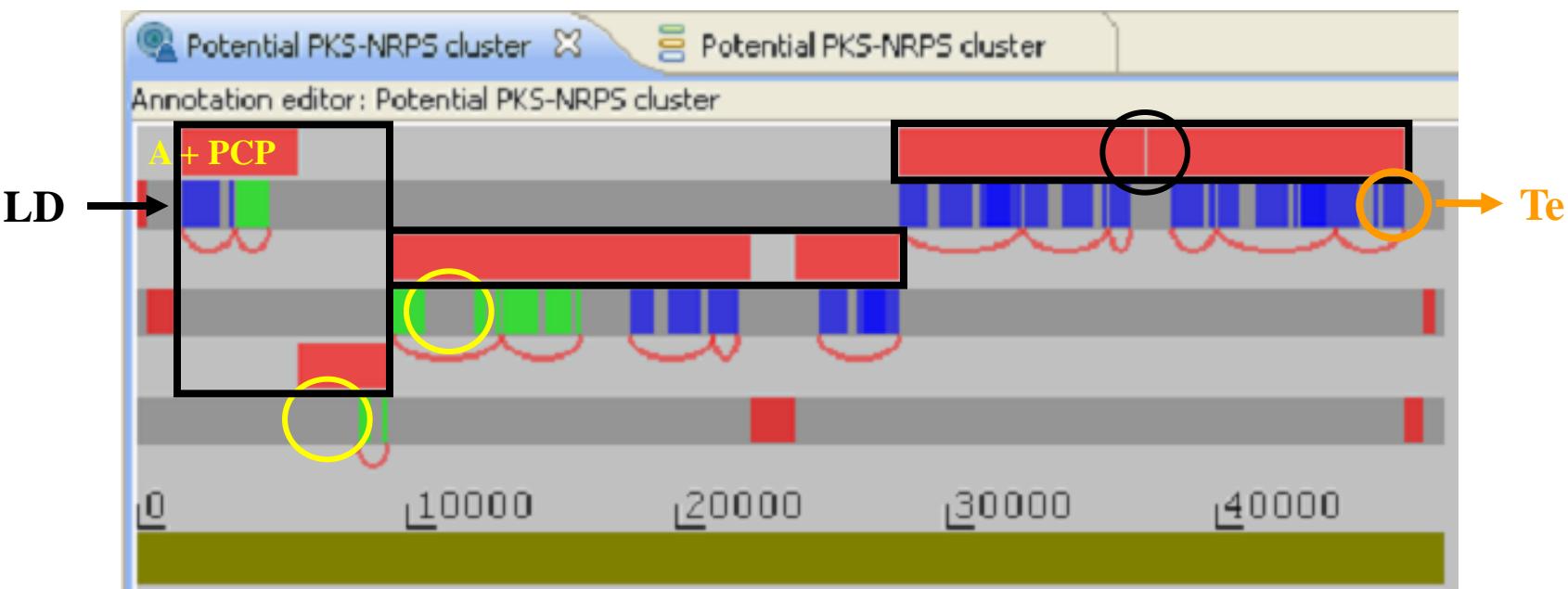
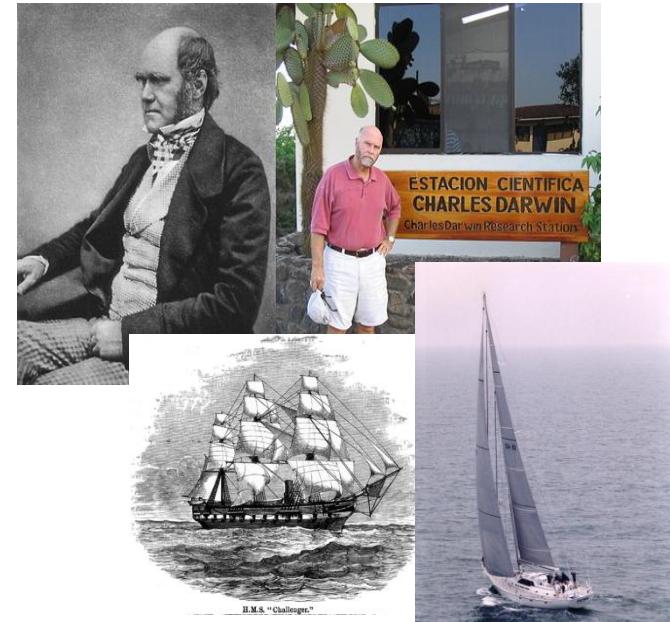
[C@ @H](C)[C@ H](O)C(C)=CC(*)C(*)C=CC=CC[C@ H](O)C(C)=C
[C@ @H](C)C(=O)CC(=O)S



- Ecopia BioSciences Inc.
Streptomyces aculeolatus and *Streptomyces* sp. Eco86
Banskota *et al.*, *J. Antibiot.*, **59**, 168–176, 2006
- Rolf Müller's group
myxobacterial species *Stigmatella aurantiaca*
Frank *et al.*, *J. Mol. Biol.*, **374**, 24-38, 2007

Sequencing of largest metagenomic dataset:

(Rusch *et al.*, PLoS Biol. 5, e77, 2007)



ClustScan downloads:

About ClustScan



ClustScan

Version: 1.0.8

Phone: +385 1 3091 900

Fax: +385 1 3091 811

E-mail: novalis@novalis.hr

web: <http://bioserv.pbf.hr/>

(c) Copyright Novalis Ltd. 2007.
All Rights Reserved.

No

ClustScan - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://reg.bioserv.pbf.hr/

Home Bookmarks The Mozilla Organiza... Latest Builds CORDIS FP6: Home ...

EBI EBI Nucleotide for

You're not currently signed in.

[HOME](#) [SIGN UP](#) [LOGIN](#)

Plug-in Data

My services

In order download ClustScan and access services, please sign in. If you don't have an account, please click sign up to make one.



CompGen program package:

- major goal of *CompGen* is structuring and maintenance of custom database of entirely novel natural products developed by *in silico* modelling of homologous recombination
- future of *CompGen* will be prediction of biological activities of *in silico* generated novel chemical entities using CADD technology

Why are Streptomyces good for recombination?

- high G+C-content - longer stretches of identity
- no *mutSL* genes - these prevent recombination between sequences with mismatches
- linear DNA allows single crossovers
 - ✓ have *recA* gene - homologous pairing
 - ✓ no *recBCD* or *addAB* – DNA helicase/endonuclease
 - ✓ 4 conserved potential DNA helicases in genomes

S. coelicolor A3(2), S. avermitilis and S. scabies

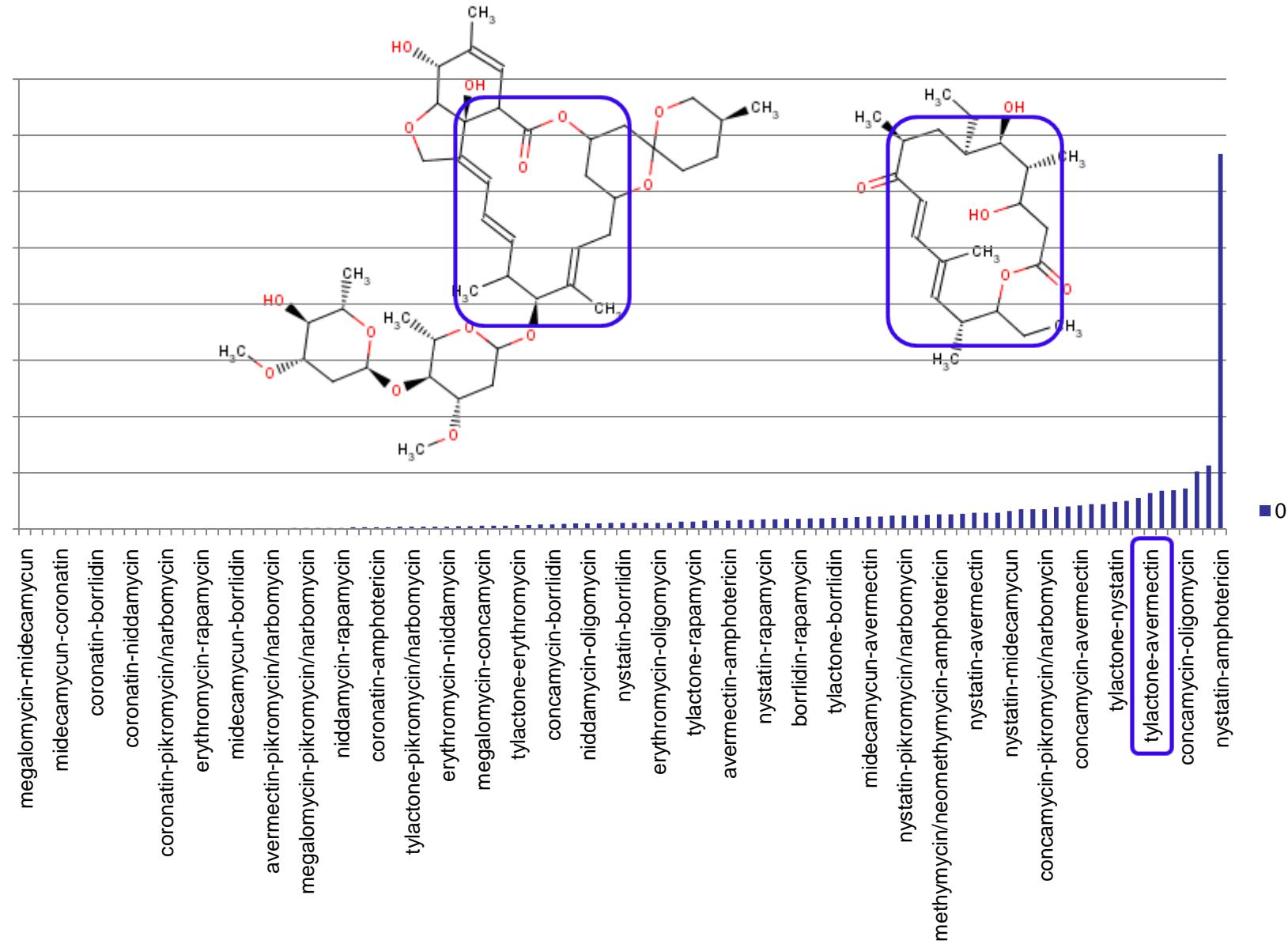
Recombination between PKS clusters:

- needs Minimal Essential Pairing Sequence (MEPS; ~ 30 bp) in region of reasonable matching (> 75% in ~ 200 bp)
- recombination program finds regions of identities and similarities
- coordinates are matched to custom database to find which domains are involved to allow product prediction

Number of recombinants:

- use parameters of 27 bp MEPS identity and 75% similarity in 200 bp
- 2445 recombination points (15 clusters)
- cca. 47 recombination points/cluster pair
- number of viable recombinants much lower, depends on additional fitness criteria imposed (2-4 rec point/cluster pair)

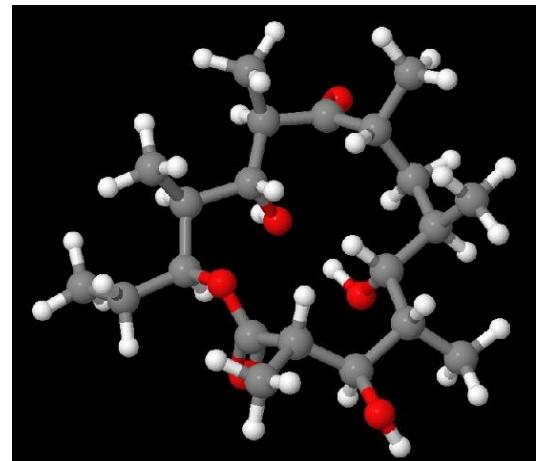
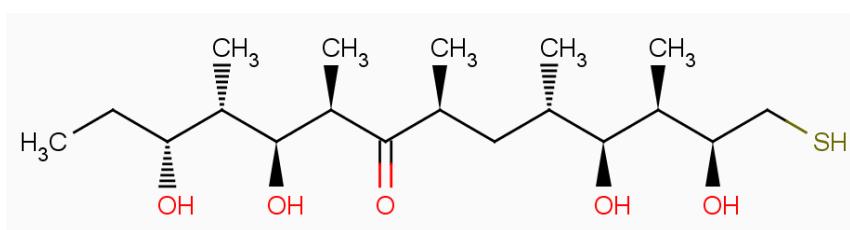
EPS



Number of EPS (Effective Pairing Sequences) in case of PKSs, correlates with overall sequence similarity, sequence length and chemical structure.

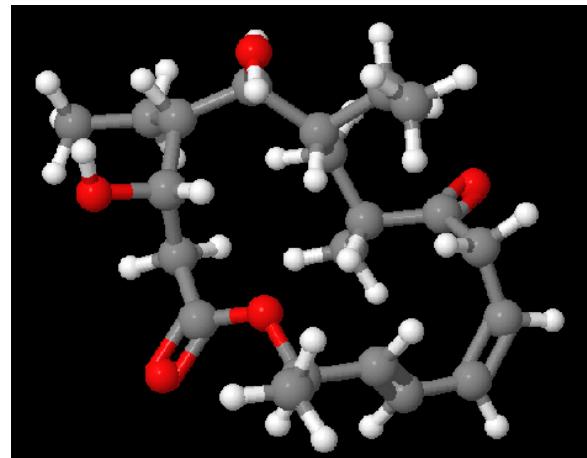
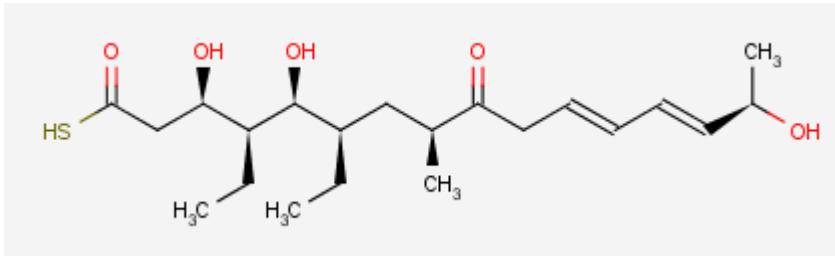
Erythromycin and niddamycin gene-clusters:

ery 14-membered macrolide



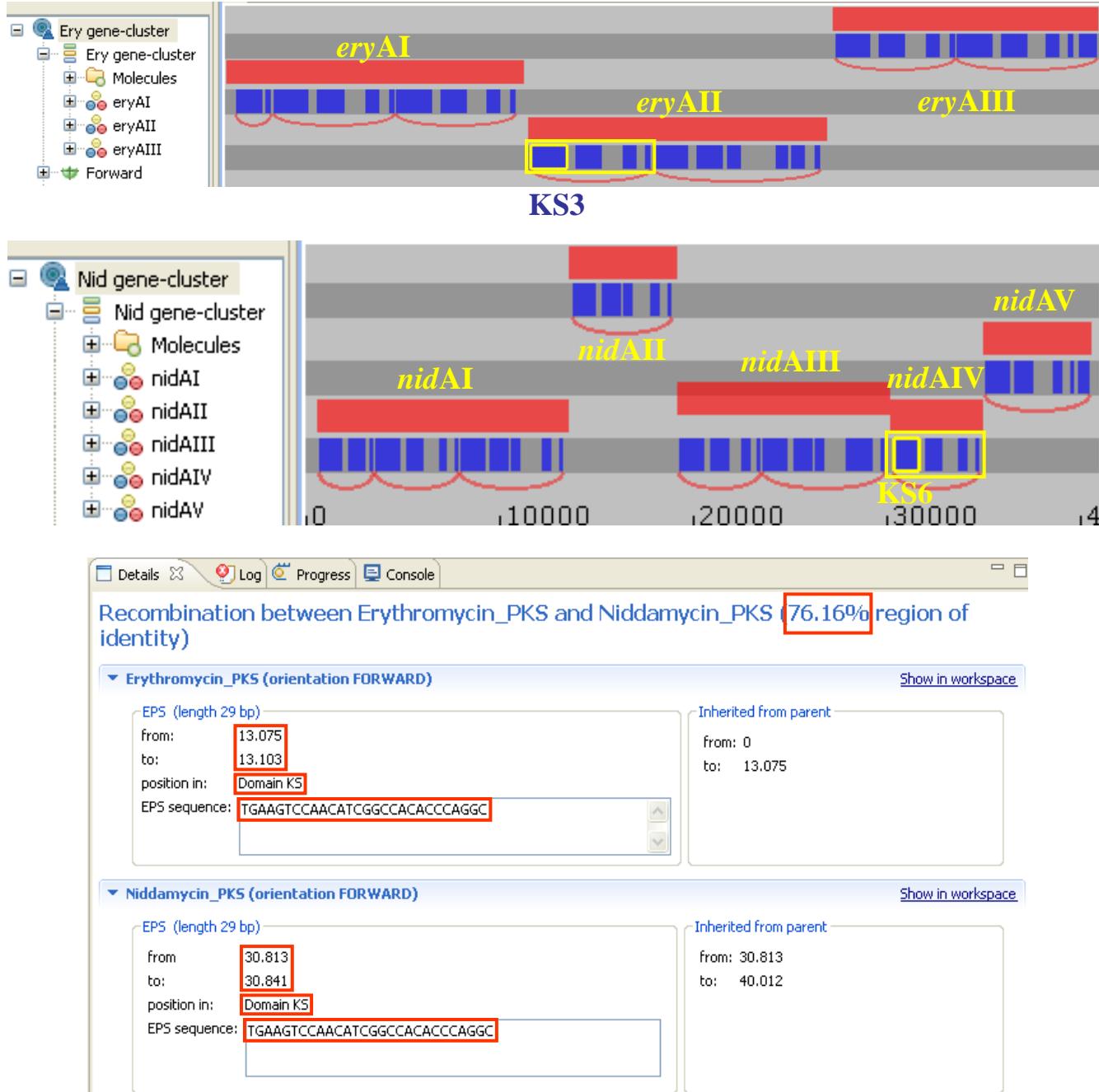
3 genes and 6 modules – 7 building blocks

nid 16-membered macrolide



5 genes and 7 modules – 8 building blocks

Recombination between PKS clusters cont':

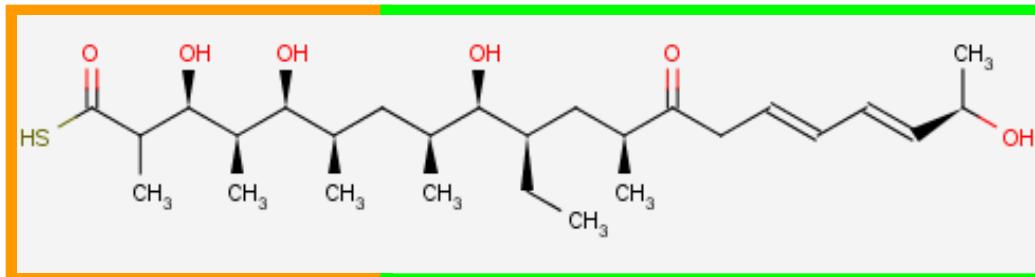
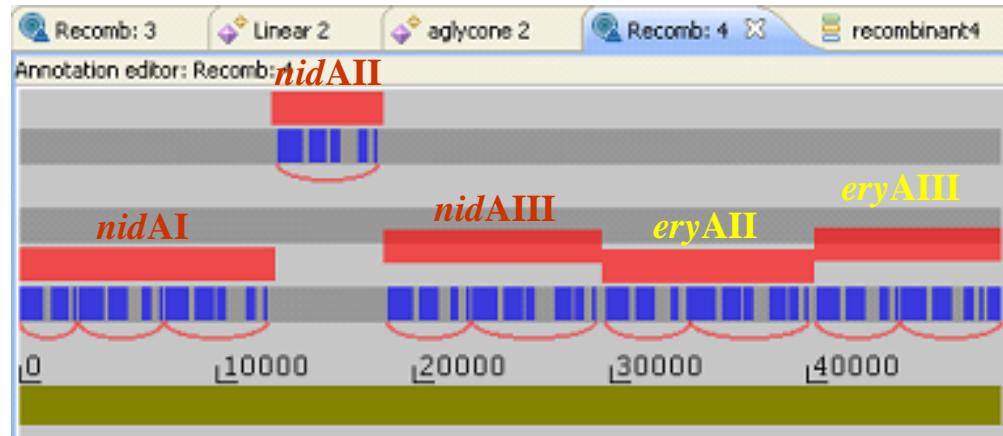


CompGen program package cont':

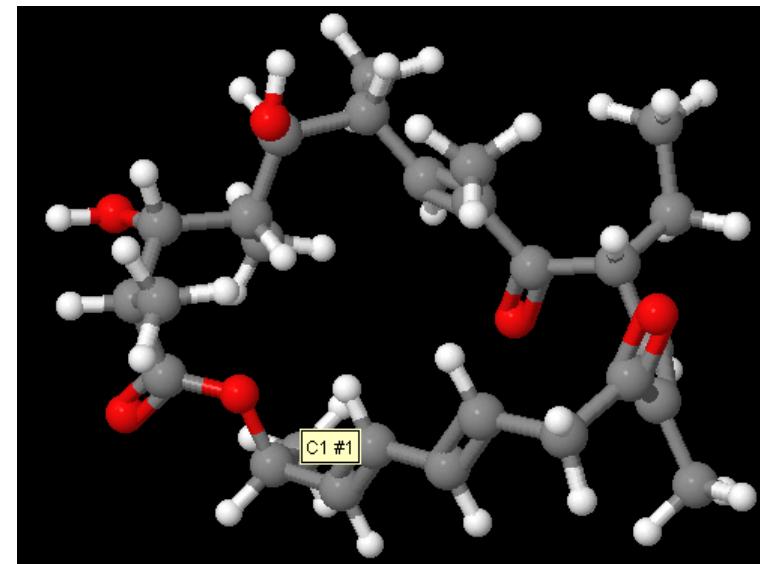
Rec-2

SMILES:

C[C@@H](O)C=CC=CCC(=O)[C@@H](C)C[C@@H](CC)[C@H](O)[C@@H](C)C[C@@H](C)[C@H](O)[C@@H](C)[C@H](O)C(C)C(=O)S



5 genes, 9 modules and
10 building blocks
18-membered macrolide

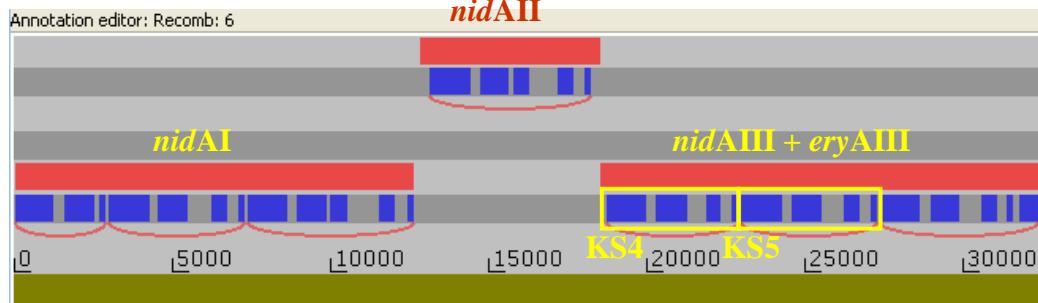


CompGen program package cont':

Rec-3

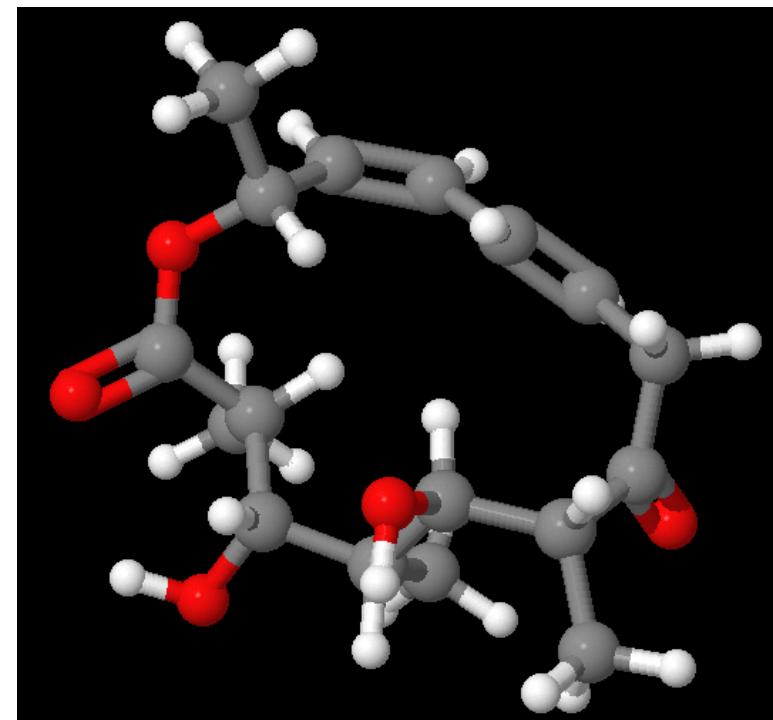
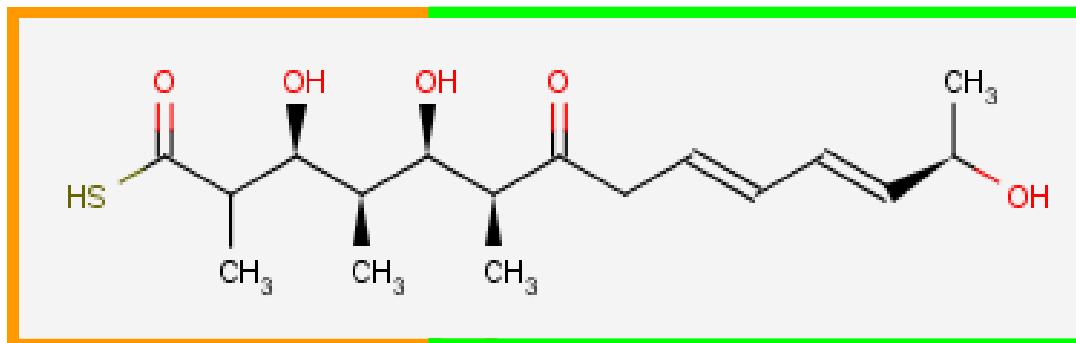
SMILES:

C[C@H](O)C=CC=CCC(=O)[C@@H](C)[C@H](O)[C@@H]2(C)[C@H](O)C(C)C(=O)S2



3 genes, 6 modules and
7 building blocks

14-membered macrolide



Summary:

- *ClustScan* prototype completed
- *CompGen* prototype completed
- reverse *CompGen* established
- when product looks promising *in silico*, "designer bug" can be created