A Folding Algorithm for Extended RNA Secondary Structures

26th TBI Winterseminar

Christian Höner zu Siederdissen, Stephan H. Bernhart,
Peter F. Stadler, and Ivo L. Hofacker

February 14, 2011

universität wien

GEN-AU
GENOME RESEARCH IN AUSTRIA

The current state of RNA secondary structure prediction

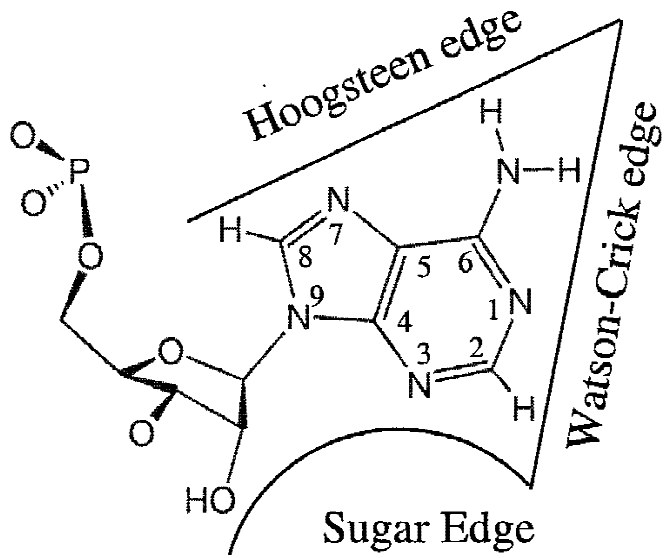
- in the ViennaRNA package:
  - only canonical structures
  - CG,GC,AU,UA,GU,UG
- in ContraFold,
- in MC-Fold:
  - all pairs, depending on the motif and a probability

The scary reality of RNA structures

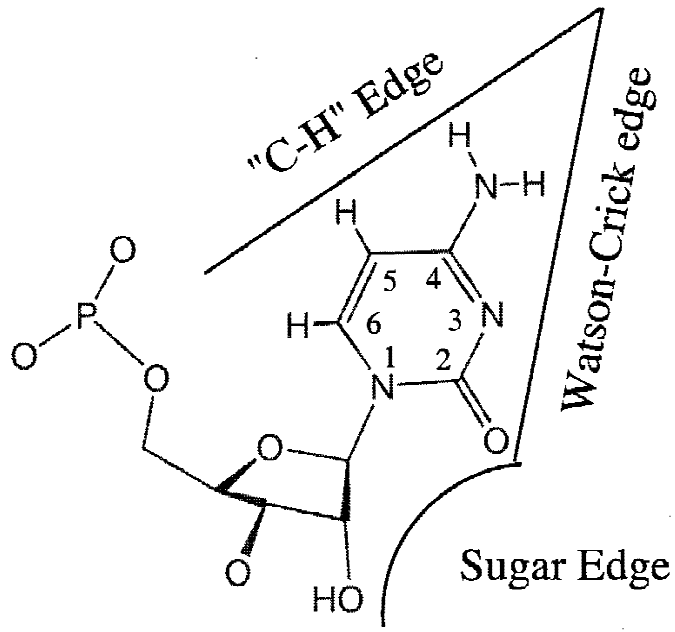
- . . . that is nice, but not true
- virtually *every* structure contains non-canonical pairs (E. Westhof!)
- all 16 possibilities [ACGU] $\times$ [ACGU]
- three different faces for each nucleotide
  - Watson-Crick
  - Sugar
  - Hoogsteen
- both *cis* and *trans* orientation
- 12 different configurations (and another one, *bif*)
- and by the way, each nucleotide can have more than one pairing partner

Let's have a picture

purine (A,G), adenine:



pyrimidine (C,U), cytosine:



(Leontis NB and Westhof E, *Geometric nomenclature and classification of RNA base pairs*, NAR, 2001)

A first try at better predictions: `MC-Fold`

- $\Psi(\text{structure}|s) =$
  $\Psi(\text{NCMs}|s) \times \Psi(\text{junctions}|\text{NCMs}) \times \Psi(\text{hinges}|\text{junctions}) \times$
  $\Psi(\text{pairs}|\text{hinges})$
- create a large database of observed motifs and frequency of occurance
- for unobserved motifs: generalize and extrapolate
- important: hinges contain non-canonical (eg. Sugar-Hoogsteen) nucleotide pairs
- the hinge type is integrated away for the final result
- runtime is somewhere in $O(15^{n/2})$ (kind of suboptimal ;-)

A small improvement: `MC-Fold-DP`

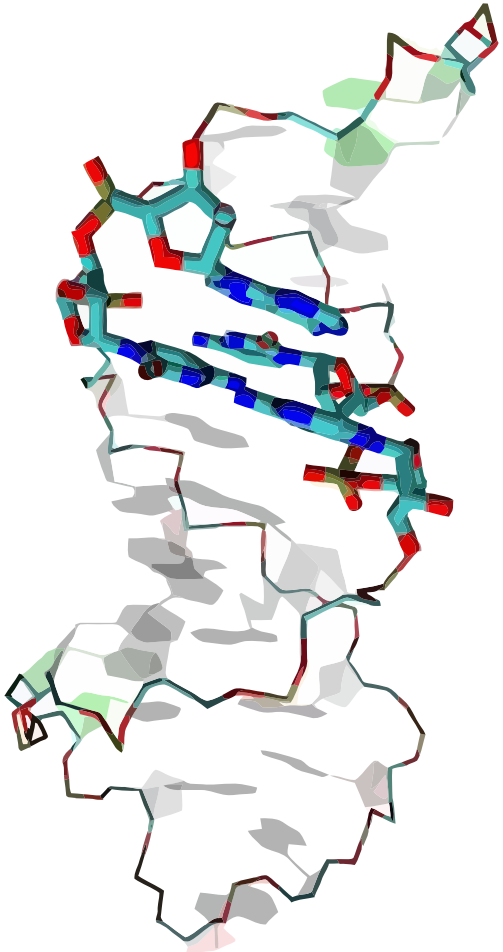
- observation: the basic structure is similar to the RNAfold recurrences
- differences: multi-branched loops and large interior loops are Nussinov-style
- for those: create a hairpin loop and fill the unpaired region with a smaller interior structure

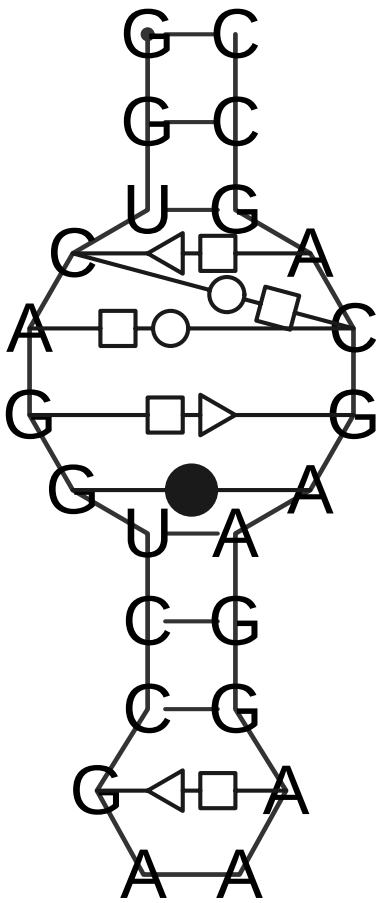
$\Rightarrow$ rewrite in a style similar to ViennaRNA (allows suboptimals, partition function, ...) leads to $O(n^3)$

- some motifs were observed only once or twice

$\Rightarrow$ added sparse data correction

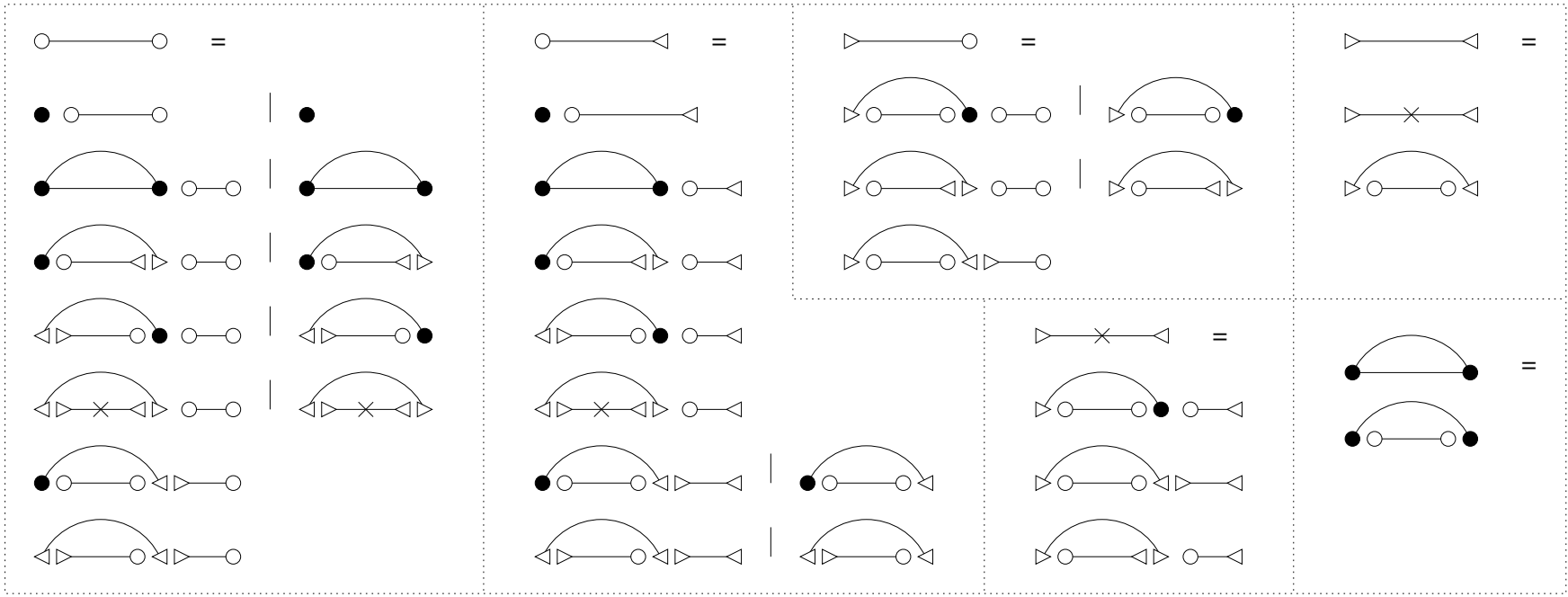
How to improve?





- make pair types explicit
  - WC
  - Hoogsteen
  - Sugar
- for each nucleotide: (allow 2-diagrams)
  - one or
  - two pairs
- train using
  - melting experiments
  - PDB (FR3D)
  - RNAstrand

2-Diagrams



(remember that each pair can be one of 12 different kinds)

Identifying problems

- melting experiments from different years (or decades) with different measured energies (small problem)
- strange PDB entries: protein interactions? errors in data? pseudoknotted structures (problem)
- many thousands of parameters, small body of evidence (big problem)
- constraints: for each database entry, the known structure should have the best energy
- parameter fitting: convex optimization problem

The prior grammar

- reduction to extended Nussinov:
  - remove stacking, consider only individual base pairs
  - loop contributions
- but keep non-canonical base pairs
- in total: ≈ 600 parameters, not $\approx 10^5$ or 10^6

For comparison: 2-stacks, eg. $\binom{C-G}{G-C}$, admit
$4 \times 4 \times 4 \times 4 \times 12 \times 12 = 36864$ parameters

Training of the prior grammar

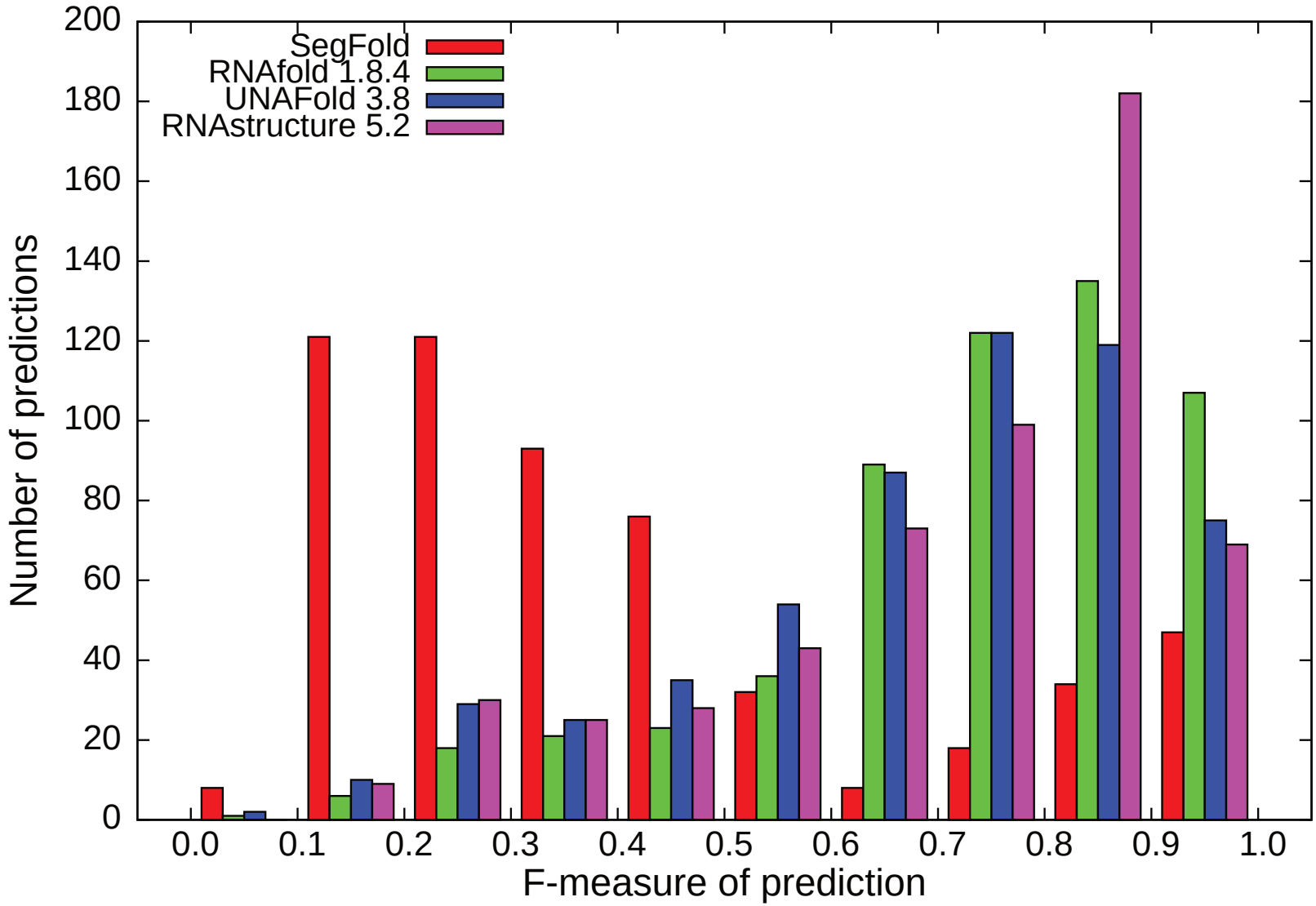
- melting energy: y, melting structural features: A
- structural constraints (known - predicted): D
  energy difference: d
- generate constraints iteratively (cf. Andronescu et al, 2007)
- destabilizing features (hairpins, bulges, interior loops): S

$$\left\| \begin{pmatrix} A & 0 \\ D & -I \end{pmatrix} \begin{pmatrix} x_{\text{cur}} \\ d_{\text{init}} \end{pmatrix} - \begin{pmatrix} y \\ d \end{pmatrix} \right\|_2$$

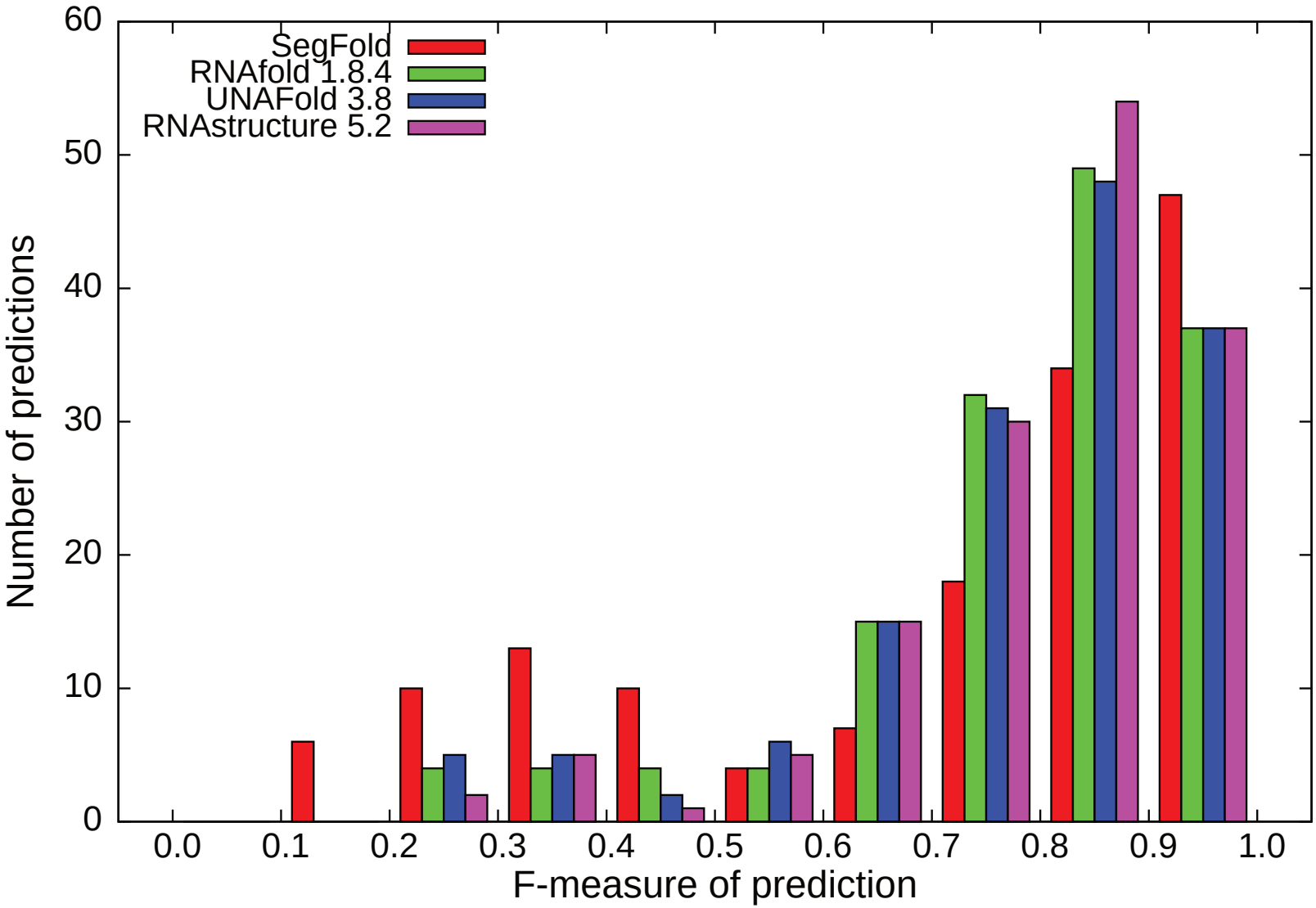
with linear constraints

$$-5 < x_j < 5, \qquad 0 < x_m, \quad m \in S, \qquad 0 < d_k$$

Results: complete set



Results: PDB only



Outlook: posterior grammar

- enable the full stacking grammar
- a full set of features
- extend the convex optimization problem to the full set of features
- train on all available databases

$\Rightarrow$ a new base algorithm for structure prediction

Thanks and other stuff

- thanks to the participants of the *Refined presentation of RNA structures* workshop
- thanks Manja for getting me off my lazy behind
  (wait for it . . .)

- we have a sensible tool for secondary structure prediction between different programs that produces good statistics
- in conjunction with ISMB/ECCB in Vienna there will be a Bioinformatics conference and a Hackathon
  `http://www.open-bio.org/wiki/BOSC_2011`
  Deadline for abstracts: 18 April 2011
  Codefest 2011: 13-14 July 2011
  BOSC 2011: 15-16 July 2011
  ISMB 2011: 17-19 July 2011