GLIMMER

Dennis Flottmann

Universität Bielefeld

Agenda

- Who invented GLIMMER?
- What is GLIMMER?
- How GLIMMER works
  - IMMs
  - ICMs
- GLIMMER live demonstration
- GLIMMER today and in comparison to other tools

Who invented GLIMMER?

- Steven L. Salzberg
- Arthur L. Delcher
- Simon Kasif
- Owen White

What is GLIMMER?

GLIMMER is a software tool, implementing a computational scoring-method to identify genes on coding regions of given DNA-sequences (procaryotic organisms)

- Desktop-application (no use of web-service necessary)
- developed under OSI License (opensource)
- customizable
- does not require many system resources (max. 50-60 MB of RAM)
- http://www.cbcb.umd.edu/software/glimmer/

How GLIMMER works

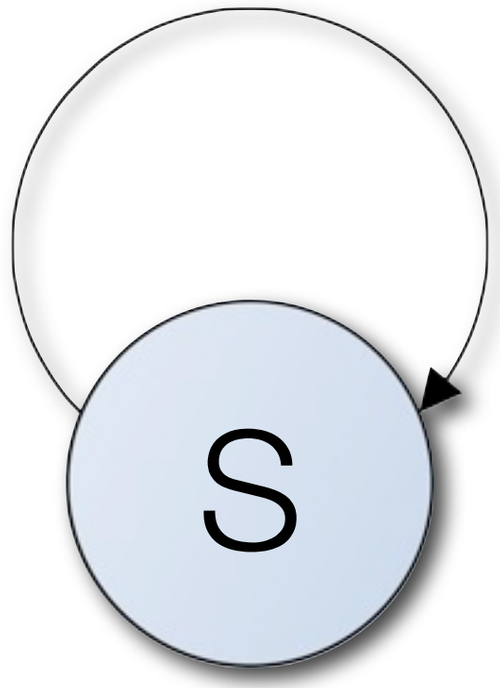
- GLIMMER calculates 7 IMM-Models (6 per reading frame + 1 non-coding regions)

- searches for all open reading frames and calculates score for all models

- orfs with adequate score will be examined for existing overlaps

- orfs with lower score then will be dismissed

Introduction to Markov chains

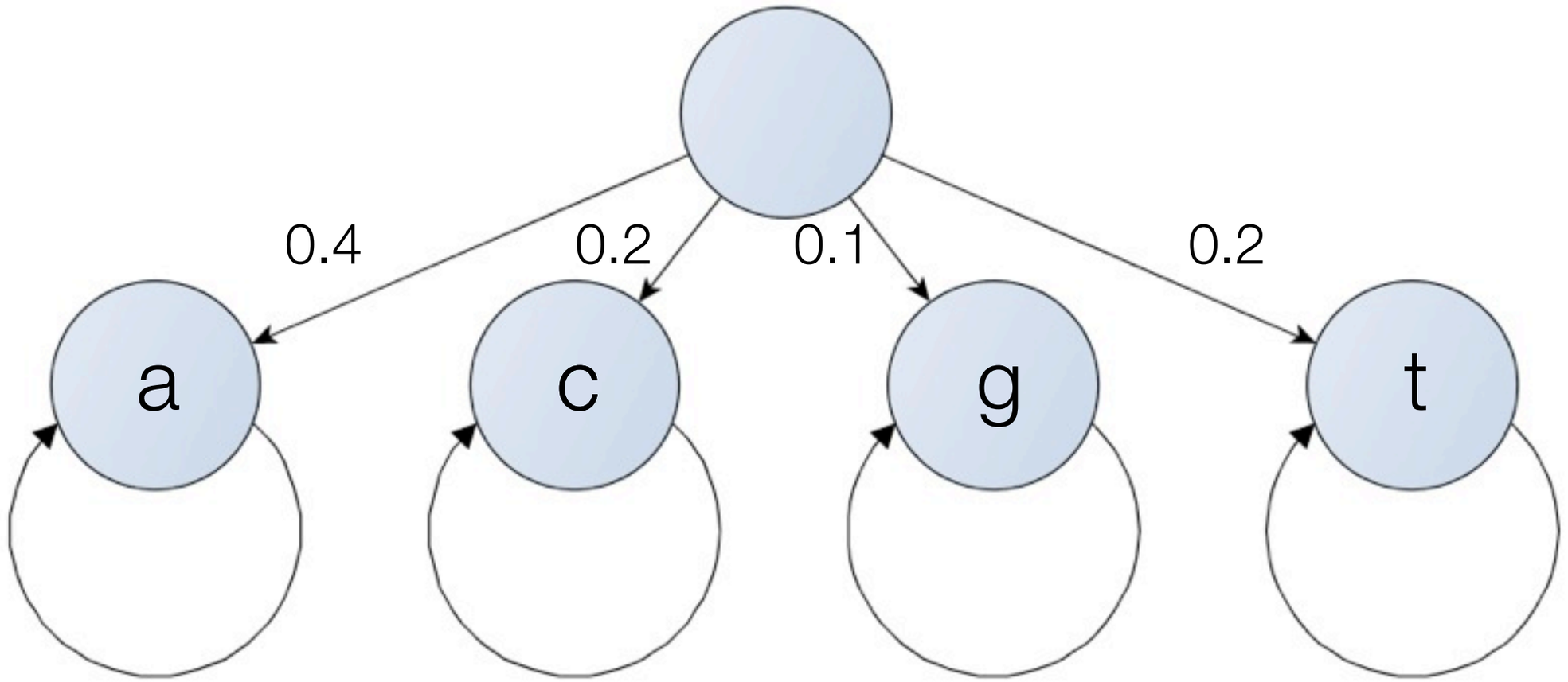


a 0.2
c 0.3
g 0.1
t 0.4

$$P(ccccc) = (0.3)^5 = 0{,}00234$$

How GLIMMER works

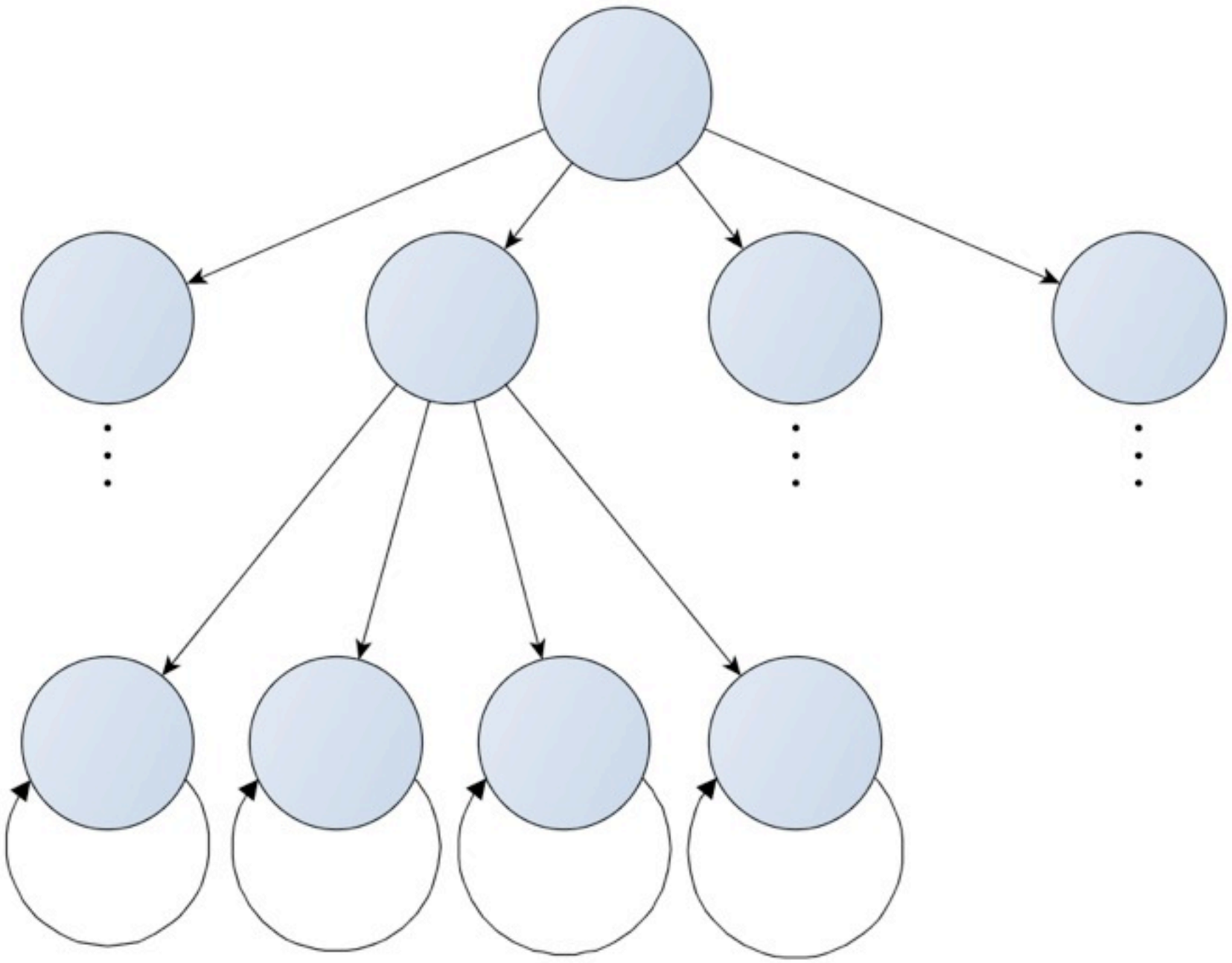
Introduction to Markov chains



a	c	g	t
a 0.7	a 0.2	a 0.3	a 0.5
c 0.1	c 0.3	c 0.3	c 0.3
g 0.1	g 0.1	g 0.3	g 0.1
t 0.1	t 0.4	t 0.1	t 0.1

How GLIMMER works

Introduction to Markov chains



How GLIMMER works

Introduction to Markov chains

- linear-combinations of Markov models
- chain of k-th order calculates the following base out of the k previous bases
- approach of Markov chains is used e.g. with GeneMark

→ GLIMMER and IMMS

How GLIMMER works

Introduction to Markov chains

- all Markov chains from 0 to 8-th order will be calculated
- chains get a weight depending on their frequency of occurence in the training-data
- if training-data is not sufficient for a higher order -> fallback to a chain of lower order

Interpolated Markov Models

Calculating the IMMs

$$P(S \mid M) = \sum_{x=1}^{n} IMM_8(S_x)$$

$$IMM_8(S_x) = \chi_8(S_{x-1}) * P_8(S_x) + (1 - \chi_8(S_{x-1})) * IMM_7(S_x)$$

$$P_i(S_x) = P(S_x \mid S_{x-i}, \ldots, S_{x-1}) = \frac{f(S_{x,i})}{\sum_{b \in \{acgt\}} f(S_{x,i}, b)}$$

Interpolated Markov Models

Calculating the weights

- weight is 1.0 if occurence of $S_{x-i} \ldots S_{x-1}$ in the training-data exceeds the threshold value (400)
- else:
  - frequency of the bases $f(S_{x,i}, b) \mid b \in \{acgt\}$ will be compared to prediction of the next shorter model IMM_{i-1}
  - if there are differences a higher weight will be given:

$$\chi_i(S_{x-1}) = \begin{cases} 0.0 & c < 0.50 \\ \frac{c}{400} \sum f(s_1 s_2 .. s_i b)_{b \in} \{acgt\} & c \geq 0.50 \end{cases}$$

How GLIMMER works

- GLIMMER calculates 7 IMM-Models (6 reading frame + 1 non-coding regions) based on training-data

- searches for all open reading frames and calculates score for all models

- orfs with adequate score will be examined for existing overlaps

- orfs with lower score will be dismissed

How GLIMMER works

- detection-rate is only ~97-98%
- much too high false-positives rate
- missing overlap-treatment causes too many unrecognized genes

How GLIMMER works

Interpolated Context Models

- ICMs: an extended version of IMMs
- the prediction of a base does not only depend on its predecessor
- the position of a base in its whole context is important!

Interpolated Context Models

Mutual Information I of two radom variables X, Y is:

$$I(X;Y) = \sum_i \sum_j P(x_j, y_j) * \log(\frac{P(x_i)P(y_j)}{P(x_i, y_j)})$$

How GLIMMER works

Interpolated Context Models

- the sequence is divided into frames of length k+1
- calculation of mutual information $I(x_1, X_{k+1}), I(X_2, X_{k+1}), ..., I(X_k, X_{k+1})$
- search maximum $I(X_j, X_{k+1})$

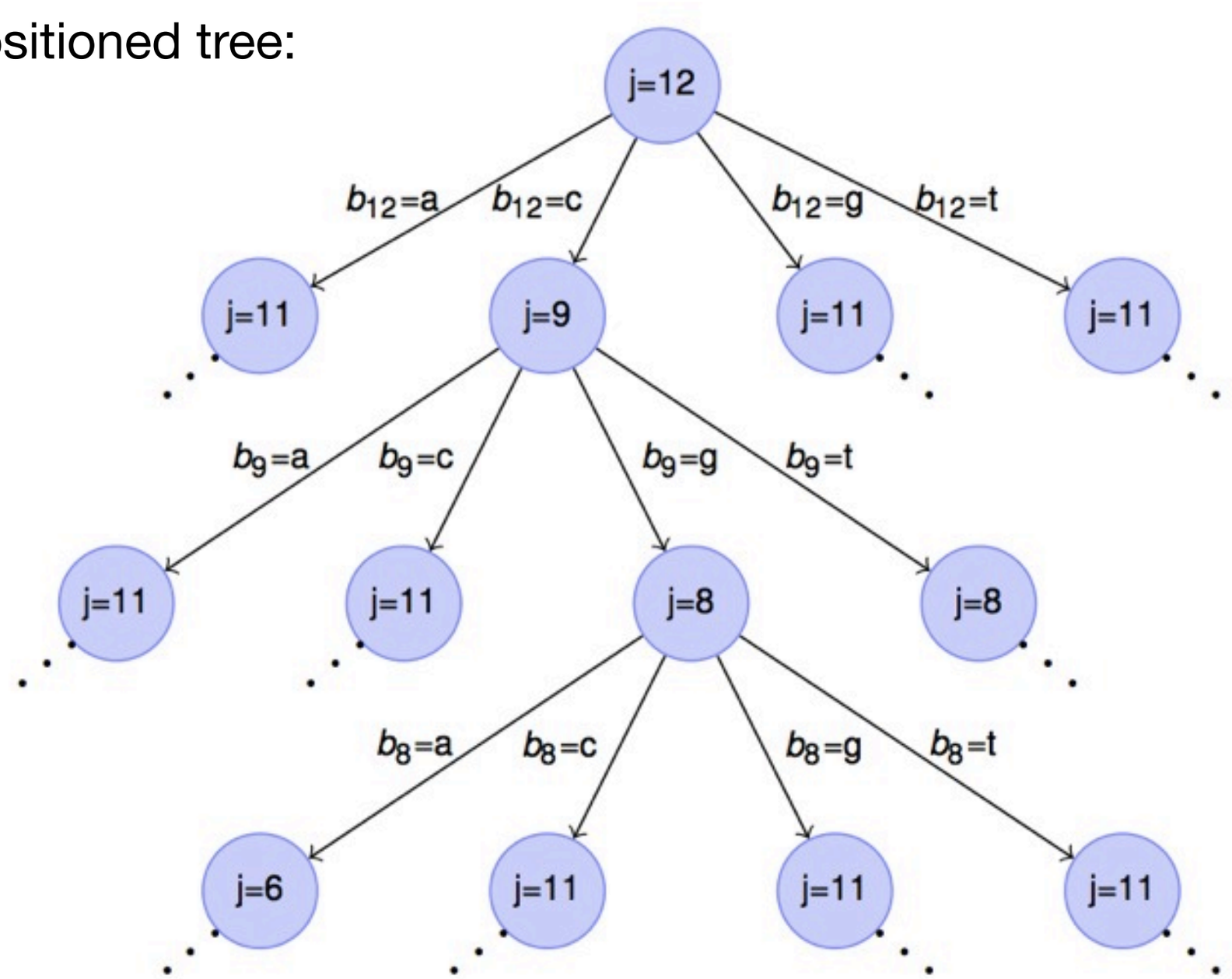
- the quantity of frames is devided into 4 sub-quantities, which are sorted according to the calculated max. position and the hereby given base
- the algorithm starts over again for each of the four sub-quantities

→ decompositioned tree

How GLIMMER works

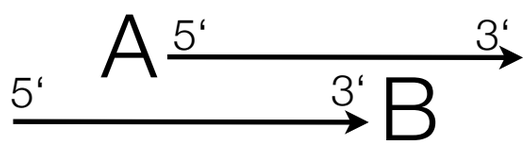
Interpolated Context Models

decompositioned tree:



Overlap treatment

- GLIMMER 2 tries to find alternative start-codon-positions
- after a gene is dismissed the recalculation of the overlaps will begin
- in the following example, gene A has a higher score at the moment:



gene B will be dismissed with high probability

How GLIMMER works

Comparison between GLIMMER 1 & 2

		GLIMMER 1.0		GLIMMER 2.0	
Organism	Genes annotated	Annotated genes found	Additional genes found	Annotated genes found	Additional genes found
H. influenzae	1738	1715 (98.7%)	234 (13.5%)	1720 (99.0%)	242 (13.9%)
M. genitalium	483	479 (99.2%)	78 (16.1%)	480 (99.4%)	82 (17.0%)
M. jannaschii	1727	1715 (99.3%)	210 (12.2%)	1721 (99.7%)	218 (12.6%)
H. pylori	1590	1545 (97.2%)	293 (18.4%)	1550 (97.5%)	322 (20.3%)
E. coli	4269	4099 (96.0%)	757 (17.7%)	4158 (97.4%)	868 (20.3%)
B. subtilis	4100	4006 (97.7%)	917 (22.4%)	4030 (98.3%)	1022 (24.9%)
A. fulgidus	2437	2385 (97.9%)	274 (11.2%)	2404 (98.6%)	341 (14.0%)
B. Burgdorferi	849	845 (99.5%)	67 (7.9%)	843 (99.3%)	62 (7.3%)
T. pallidum	1039	1012 (97.4%)	180 (17.3%)	1014 (97.6 %)	250 (24.1%)
T. maritima	1877	1849 (98.5%)	190 (10.1%)	[illegible]	208 (11.1%)

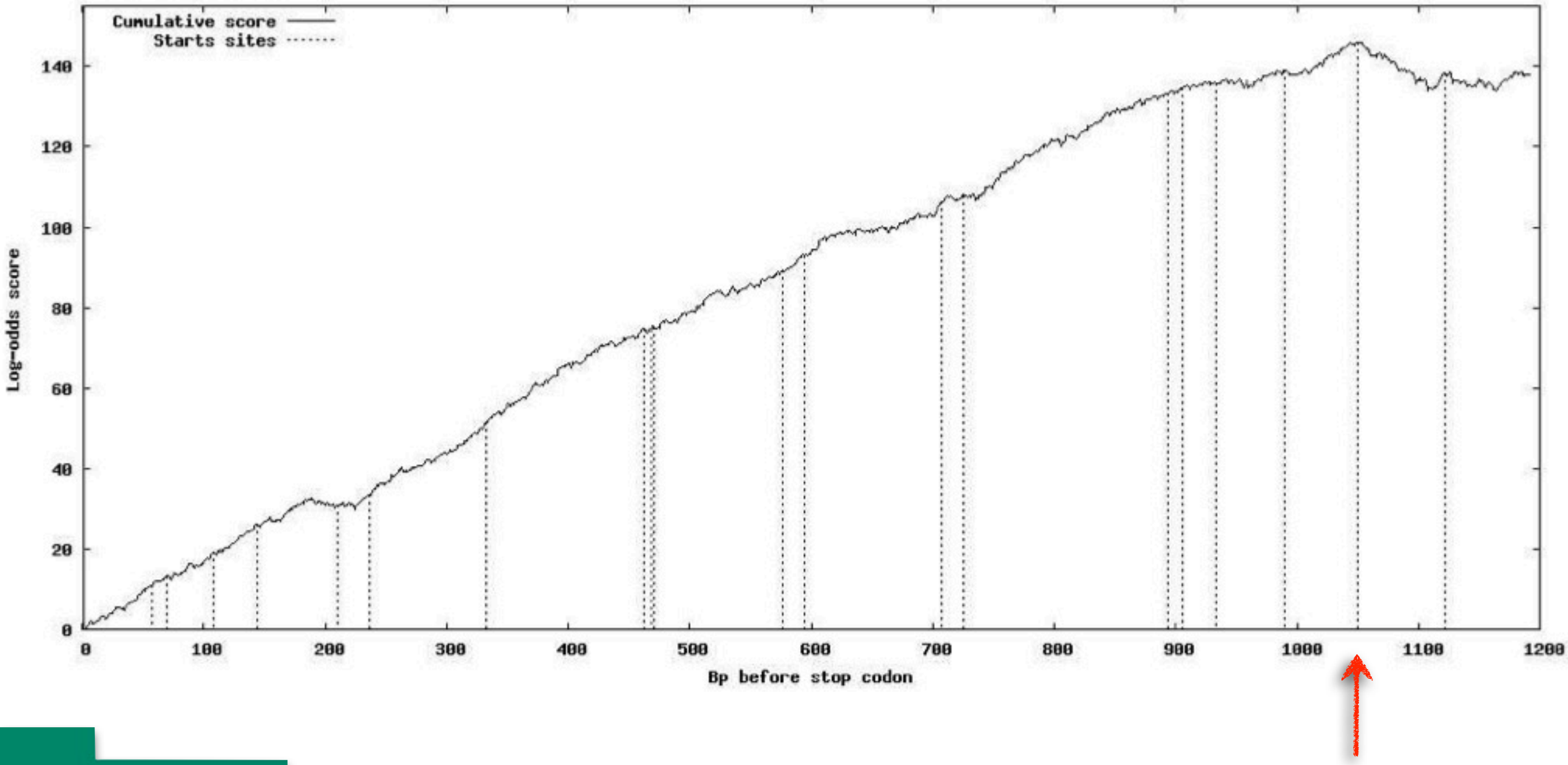
0.5 % increased accuracy

GLIMMER today and in comparison to other tools

- calculates the score backwards beginning with the stop-codon
  - because IMMs are only trained for genes (transition from coding to non-coding of a context-frame result in low scores)
  - score is added up (reaching the correct start-codon results in a maximum score)
- GLIMMER 1/2 preferred longer orfs; GLIMMER3 prefers higher scores

GLIMMER today and in comparison to other tools

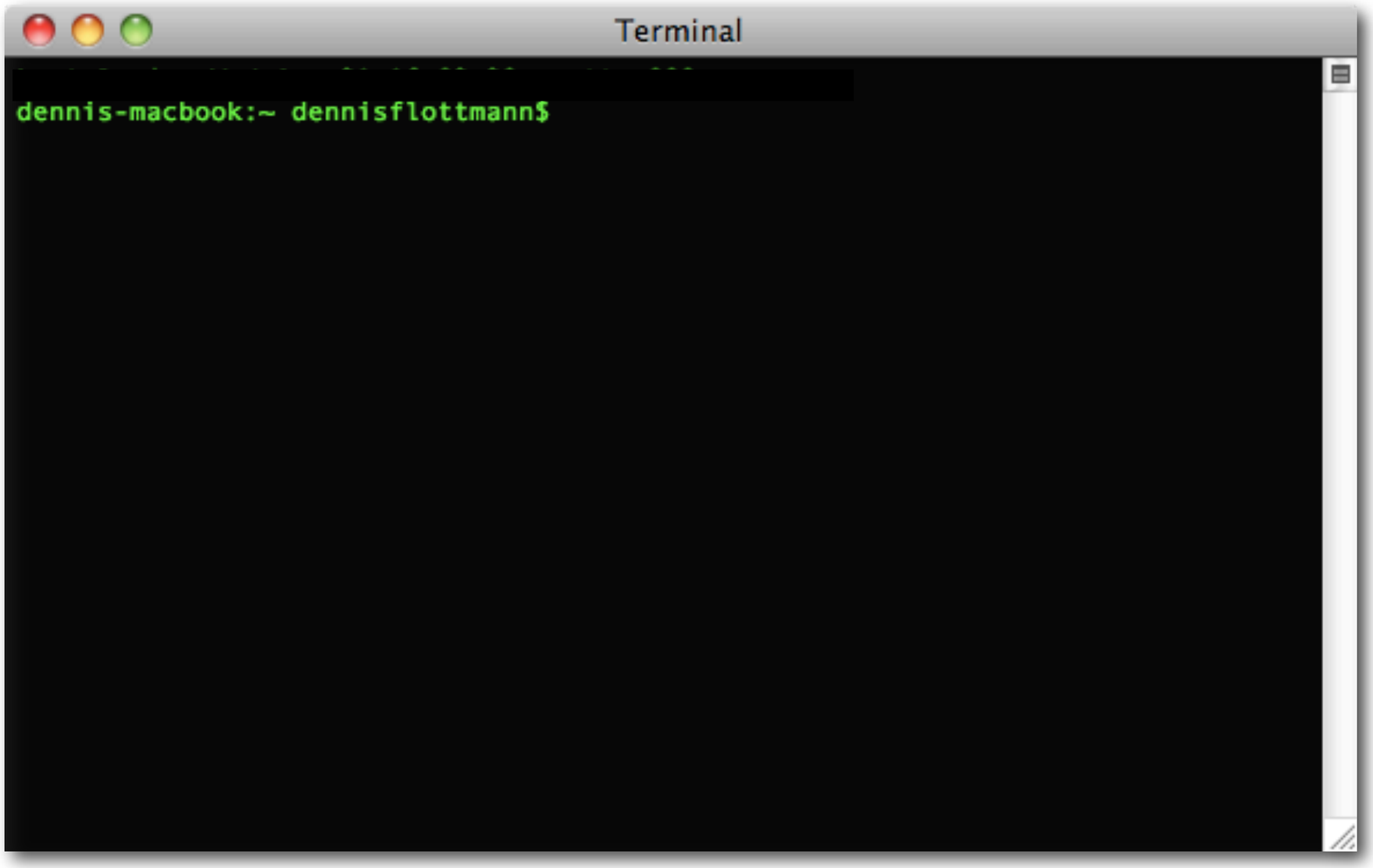
GLIMMER 3



Other improvements

- Ribosomal binding sites can give a strong hint for the correct start-codon
  - ELPH searches for motifs in the quantity of sequences
  - GLIMMER uses created PWM to score potential RBS
- Overlaps
  - GLIMMER3 calculates **every** possible orf between start- & stop-codon
  - a dynamic algorithm tries to combine a quantity of orfs into a valid sequence with maximum total-score and a minimum of overlaps
- improved long-orf training
  - GLIMMER2 chooses orfs > 500bp
  - GLIMMER 3 determines threshold independently as long as there are no overlaps

GLIMMER live presentation



Other tools

- **GeneMark** (Borodovsky et al. 1993)
  - GeneMark.hmm
  - GeneMarkS
    - also eucaryotic versions available
- **EasyGene** (Larsen et al. 2003)

Comparison: GLIMMER3 vs. ...

Genome		vs. GeneMark.hmm			vs. EasyGene 1.2			vs. GeneMarkS		
Organism	# Genes	3‘ Match	5‘ & 3‘	Extra	3‘ Match	5‘ & 3‘	Extra	3‘ Match	5‘ & 3‘	Extra
A. fulgidus	1165	+4	-20	-86	+5	-25	+119	0	+2	-71
B anthracis	3132	-2	-48	-134	+13	-63	+175	+1	+412	-142
B. subtilis	1576	+2	+280	+87	+15	-10	+536	-5	-39	+193
C. tepidum	1292	+1	+21	+19	+10	+9	+182	+1	-14	+29
C. perfringens	1504	-2	+177	-120	-2	-8	-21	-3	-14	-139
E. coli	3603	-25	+18	+188	+60	+44	+407	-25	-29	+190
G. sulfurreducens	2351	+13	+215	+34	+5	-1	+60	+14	+41	+66
H. pylori	915	-1	-3	-55	+4	-6	+148	-1	-8	-41
P. flourescens	4535	+17	+288	+59	NA	NA	NA	+17	+479	+46
R. solanaccearum	2512	+7	+183	+225	+11	+48	+193	-3	+160	+190
S. epidermidis	1650	+3	-32	-40	NA	NA	NA	+6	+204	-64
T. pallidum	575	+2	-8	+94	+8	-8	+176	-2	-18	+90
Averages:		+2	+89	+23	+13	-2	+198	+1	+98	+29

Customization

Abstract

- GLIMMER is a gene-finding tool that recognizes 97-98% of all genes in a prokaryotic genom

- also an eucaryotic version is available (GLIMMERHMM)

- by chosing certain training-data also other tasks can be realized

- online-version available

Questions?

Thank you for your attention!