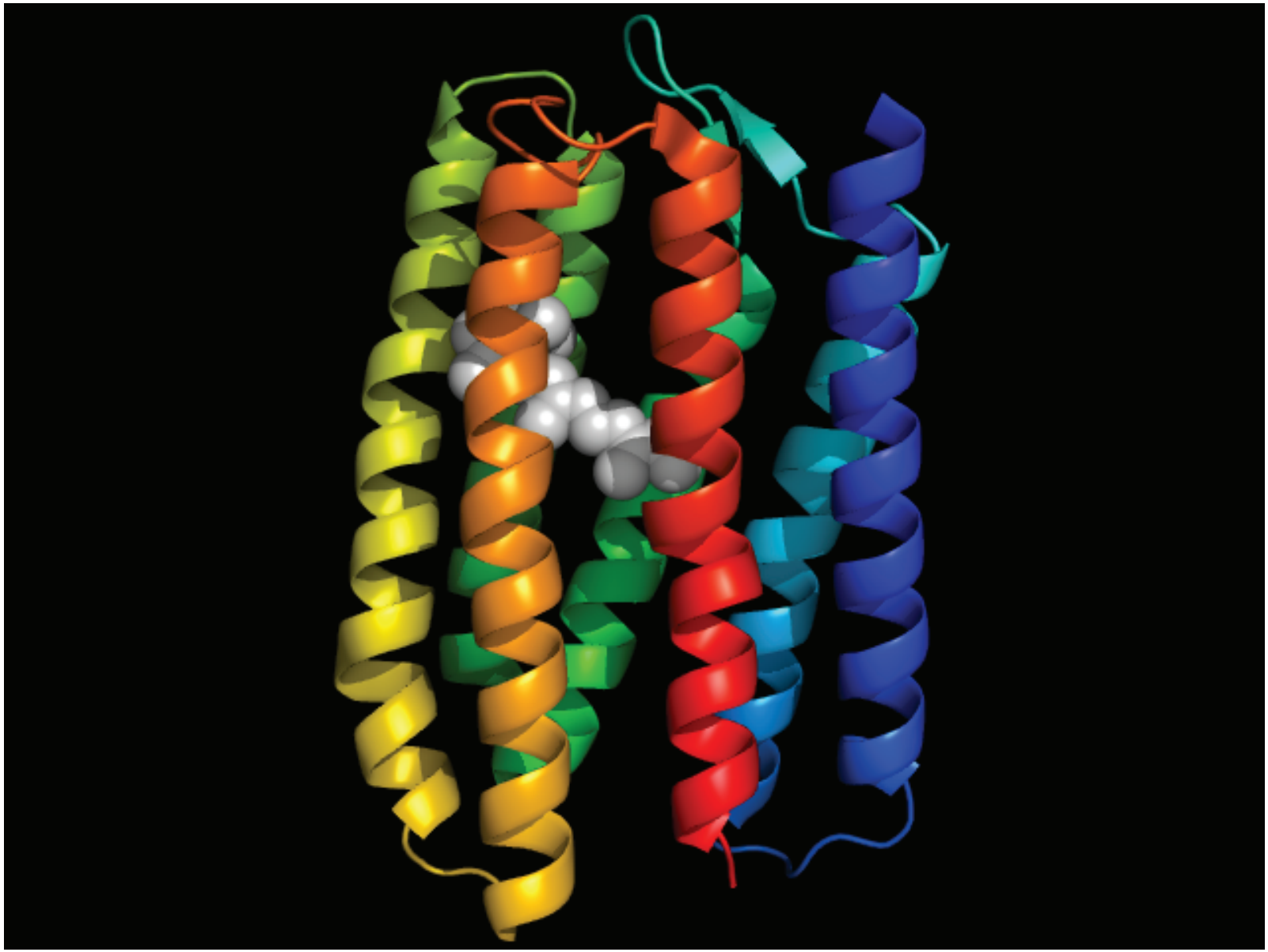
Is it worth to parameterize sequence alignment with an explicit **evolutionary model** ?

Sean Eddy & E.R.



Channelrhodopsin-1

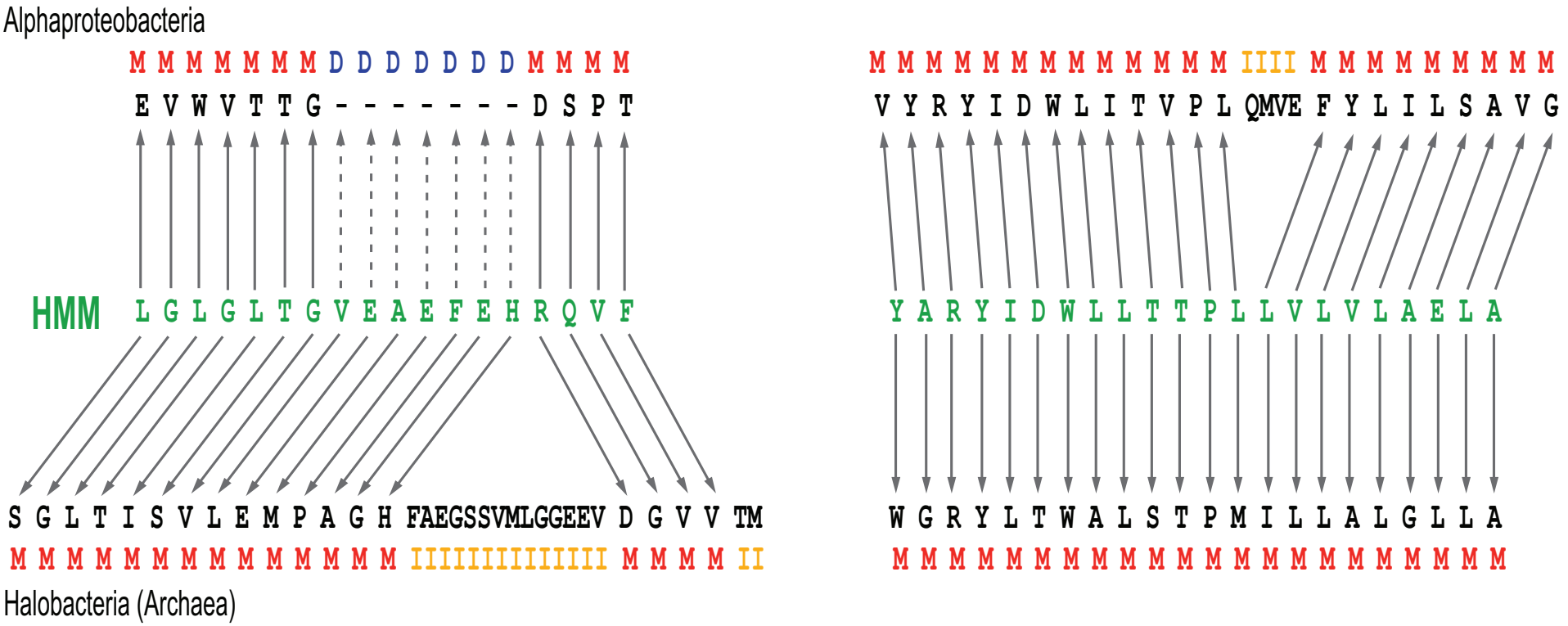
adapted from www.calvin.edu

Bacterial Rhodopsins

```
BACS2_HALSA   .TWFWVGAVGMLAGTVLPI..RD  CIRHP  SHRRYDLVLAGITGLAAIAYTTMG  LGITATTVGD.................RTVY..  LARYIDWLVTTPL....IVLYLAMLA  RPG.................
BACS2_NATPH   TTLFWLGAIGMLVGTLAFAWAGR  DAGSG  E.RRYYVTLVGISGIAAVAYVVMA  LGVGWVPVAE.................RTVF..  APRYIDWILTTPL....IVYFLGLLA  GLD.................
BACS2_HALVA   TTWFTLGLLGELLGTAVLAY.GY  TLVPE  ETRKRYLLLIAIPGIAIVAYALMA  LGFGSIQSEG.................HAVY..  VVRYVDWLLTTPL....NVWFLALLA  GAS.................
BACS1_HALSA   ATAYLGGAVALIVGVAFVWLLYR  SLDGS  PHQSALAPLAIIPVFAGLSYVGMA  YDIGTVIVNG.................NQIV..  GLRYIDWLVTTPI....LVGYVGYAA  GAS.................
C7P1Y4_HALMD  TTVYGLTAVVYAVALVVLWGWLR  QV.SP  EHRRFCTPIVLVVALAGVASAVVA  AGVGTITVNG.................SEVV..  VPLFVESMIAYGV....LYAVMARLA  DVE.................
D3SUL9_NATMM  FVLLVVSSIVFISAAAIFVGYSR  TLPDG  PNQYGYAAAVA.AGSMGLAYVVMA  LVNGISG.....................ADTD..  LFRFLGYTAMWTV....IVLVVCSVA  GVD.................
BACH_NATPH    ASSLYINIALAGLSILLFVFMTR  GLDDP  RAKLIAVSTILVPVVSIASYTGLA  SGLTISVLEMPAGHFAEGSSVMLGGEEVDGVVTM  WGRYLTWALSTPM....ILLALGLLA  GSN.................
BACR_HALAR    AIWLWLGTAGMFLGMLYFIARGW  GETDS  RRQKFYIATILITAIAFVNYLAMA  LGFGLTIVEFAGEE.............HPIY..  WARYSDWLFTTPL....LLYDLGLLA  GAD.................
BACR_HALSA    WIWLALGTALMGLGTLYFLVKGM  GVSDP  DAKKFYAITTLVPAIAFTMYLSML  LGYGLTMVPFGGEQ.............NPIY..  WARYADWLFTTPL....LLLDLALLV  DAD.................
BACR1_HALSS   TLWLGIGTLLMLIGTFYFIVKGW  GVTDK  EAREYYSITILVPGIASAAYLSMF  FGIGLTEVQVGSEM.............LDIY..  YARYADWLFTTPL....LLLDLALLA  KVD.................
B6BSG6_9PROT  GISFWVISMGMLAATAFFFMETG  NVAAG  W.RTSVIVAGLVTGIAFIHYMYMR  EVWVTTG.....................DSPT..  VYRYIDWLITVPLQMVEFYLILSAVG  KAN.................
C4YF64_CANAW  WAAFSVFLLLTIIHLLLFLYGNF  R.KPG  VKNSLLVIPLFTNAVFSVFYFTYA  SNLGYAWQAVEFQH.........AGTGLRQIF..  YAKFIAWFVGWPA.....VLALFEIV  TST.VLDRIEENPNIFKKFFLI
B9W6Y7_CANDC  WAVFSVFALFAIVHGFIYSFTDV  R.KSG  LKRALLTIPLFNSAVFAFAYYTYA  SNLGYTWILAEFNH.........AGTGFRQIF..  YAKFVAWFLGWPL.....VLAIFQIV  TNT.SFTTTEDESDLLKKFISL
A3LUH9_PICST  WALFSVFSLFAVVHAFVYGFTSS  E.KKS  LKKTLLVIPLFINAVMAYTYFTYA  SNLGWTSTPTEFQH.....VTTSEDLDVRQIF..  YVKWVGYFLTWPL.....VLTIIEVT  TQS...TDFFEEGDILTKFFSL
B5RTR5_DEBHA  WAVFSIFATLAVVHAFVFSFTSS  R.THR  LKKILFIVPLFTNAIMAYCYFTYA  ANLGWTSTRVEFNH.....VSTNRLLGVRQVF..  YVKYIGWFLAWPF.....VLFAIEVA  THTLESTNLADGGETVTGILSL
C5E3Q5_LACTC  WAVFSVFGLVSLIYAALFVVFEH  R.GTK  IHRYAVAGPLSISLVLAFSYFTMA  SNLGWTAVQAEFNN..LTTPNQSEVPGIRQIF..  YAKYVAWFLTWPA.....LLYLTELT  GVV.TRDSSNILGPRPWSFYDL
C5DYF7_ZYGRC  WTVTAIFGLLAVVYVLLFFVTQV  RNGSG  LSRYSLAAPFLIAFFEFFAYFTYA  SNLGWTGTNAEFHHISVSKPVTGESPGIRQVF..  YCKYIAWFLSWPI.....VLFLQDLA  ALS....TIKRDALGSASVLDL

HMM           WTVFSVGALLALVGTLLFFVTAR  RVKDG  EKRKLLVILLLIPAIAAVAYVTMA  LGLGLTGVEAEFEH--------------RQVF--  YARYIDWLLTTPL----LVLVLAELA  GAD------------------

HRTSAWLLAADVFVIAAGIAAAL  T..T  GVQ.....RWLFFAVGAAGYAALLYGLL  .GTLPRALGDDPR  VR..SLFVTLRNITVVL...WTLYPVVWLL  SPAGIGILQ  TEMYTIVVVYLDFISKVAFVAFAVLGADA  VSRLV
SREFGIVITLNTVVMLAGFAGAM  V..P  GIE.....RYALFGMGAVAFLGLVYYLV  .GPMTESASQRSS  GIK.SLYVRLRNLTVIL...WAIYPFIWLL  GPPGVALLT  PTVDVALIVYLDLVTKVGFGFIALDAAAT  LRAEH
REDTVKLVVLQALTIVFGFAGAV  T..P  SPV.....SYALFAVGGALFGGVIYLLY  .RNIAVAAKSTLS  DIEVSLYRTLRNFVVVL...WLVYPVVWLL  GAAGVGLMD  VETATLVVVYLDVVTKVGFGVIALLAMID  LGSAG
RRSIIGVMVADALMIAVGAGAVV  T..D  GTL.....KWALFGVSSIFHLSLFAYLY  .VIFPRVVPDVPE  QI..GLFNLLKNHIGLL...WLAYPLVWLF  GPAGIGEAT  AAGVALTYVFLDVLAKVPYVYFFYARRRV  FMHSE
GRALAAIVLTPVVQRIAFEVAAV  S..G  GIV.....ALIGLVVVVGGHLAIAAYLL  .GPVWTQTRGVPE  QRR.LLHWKARNLVLFLIGMLIAYAVIALF  GVF.....D  AFVSLAISQYMAVLIRVGFAGFLLANLDA  VGSAS
RRLTLFLFAAVLGRLWITLGSWF  V..D  GTL.....ALVATLGTFAALGFGLYLLF  .GPFTRAAAALES  ERR.LLFSKLKYLIVLG...WVGL.VATGI  MAQGAGLAD  DFVGQLVVIYVEVILILGFGAIVVRSRTA  LSQTA
ATKLFTAITFDIAMCVTGLAAAL  TTSS  HLM.....RWFWYAISCACFLVVLYILL  .VEWAQDAKAAGT  A...DMFNTLKLLTVVM...WLGYPIVWAL  GVEGIAVLP  VGVTSWGYSFLDIVAKYIFAFLLLNYLTS  NESVV
RNTITSLVSLDVLMIGTGLVATL  SPGS  GVLSAGAERLVWWGISTAFLLVLLYFLF  .SSLSGRVADLPS  DTR.STFKTLRNLVTVV...WLVYPVWWLI  GTEGIGLVG  IGIETAGFMVIDLTAKVGFGIILLRSHGV  LDGAA
QGTILALVGADGIMIGTGLVGAL  T.KV  YSY.....RFVWWAISTAAMLYILYVLF  .FGFTSKAESMRP  EVA.STFKVLRNVTVVL...WSAYPVVWLI  GSEGAGIVP  LNIETLLFMVLDVSAKVGFGLILLRSRAI  FGEAE
RVSIGTLVGVDALMIVTGLVGAL  S.HT  PLA.....RYTWWLFSTICMIVVLYFLA  .TSLRAAAKERGP  EVA.STFNTLTALVLVL...WTAYPILWII  GTEGAGVVG  LGIETLLFMVLDVTAKVGFGFILLRSRAI  LGDTE
SGMFWRLLLGSVVMLVGGYLGEA  ...G  YIN.....ATLGFIIGMAGWVYILYEVF  SGEAGKAAAKSGN  KALVTAFGAMRMIVTVG...WAIYPLGYVF  GYLTGGV.D  AESLNVVYNLADFVNKIAFGLVIWAAATS  SSGKR
FQTWLVKFIFVEIYVLGLLIGSI  I..F  STY.....KFGYFTFAVFFQLLLMVWVG  .RDLHRSFKSPSH  S...NIANFFLIFFYLV...WILYPVAWGL  SEGGNVI.Q  PDSEAVFYGILDLITFGLMPTILIFFAIK  GCDEE
FEALFTRVLAIEVFVLGLLIGAL  I..E  STY.....KWGYFTFAVVFQLFAIYLVI  .NDVVVSFGSSSH  S...VFGNALILAFVVV...WILYPVAWGL  SEGGNVI.Q  PDSEAVFYGILDLITFGVIPIILTWIAIN  NVDEE
FSRLFAKILATEVFVIGLLIGAL  I..E  STY.....KWGYFTFSVTAQLFAEIYIF  .VNVMTAWRQSTQ  ....KLGLILVLCQLVI...WILYPIAWGL  SEGGNKI.Q  PDSEAAFYGVLDFFTFFFIPVGLTWLAIN  NVDEE
LSGLIVKTFATEIYVLGLLIGIL  I..P  SSY.....RWGYFTFAVSAQLFAMSLIL  .VSMFSAAKSVHT  N...KAAIIFIAFQLLV...WILYPICWGL  SEGGNRI.Q  PDSEAVFYGILDLITFSFVPIILTWINAS  GVDED
VHGLFLQICGSWFFIIGLLVGSL  I..H  SSY.....KWGYWTMAAFAQLLVTYLIF  ....KHQLVDLTI  S...GIKLVLLVFTHVC...IYLYLVAWGL  SDGGNVI.T  VDSSHVFFGILDLLIFVLVPALLVATATS  SGVMP
IHSLLVQIFGHYFWVIALLVGAL  I..P  STY.....RWGYWTIGAFTMLVTEGLVL  ....QRQVQALRT  R...GIYLILLMFMCLI...VWCYFIAWAV  SEGGNKI.Q  PDSEAVFYGILDLVVFAIYPSILVWIITV  RGEWP

RRTLLVLVLADVVMIVGGLVGAL  I--E  STY-----RWVYFTISVAAQLVLLYLLL  -GELARAAKSLSS  EI--SLFNTLRNLVVVL---WLLYPVAWLL  GEEGNGI-Q  ADSEAVVYGILDLVAKVGFGLILLASATS  NES--
```

static (fixed in time) HMM

Bacterial Rhodopsins



→ **M** an evolved ancestral residue (a substitution)

→ **D** a deleted ancestral residue

→ **I** an insertion relative to the ancestral sequence

evolutionary (time-dependent) HMM

Homology is an evolutionary question

Homology detection is hypothesis testing

`Forward score`

$$F = \log \frac{P(\mathrm{s} \mid \mathbf{H})}{P(\mathrm{s} \mid \mathrm{R})}$$

`Posterior of H given s`

$$P(\mathbf{H} \mid \mathrm{s}) = \frac{\mathrm{e}^{\mathrm{F}+\rho}}{1+\mathrm{e}^{\mathrm{F}+\rho}}$$

Evolutionary distance is a nuisance parameter in $P(\mathrm{s} \mid \mathbf{H})$

Current approaches assume (implicitly or explicitly)

a fixed evolutionary distance

An explicit time-parameterization

allows to

Integrate over Evolutionary Distance

`Homology detection`

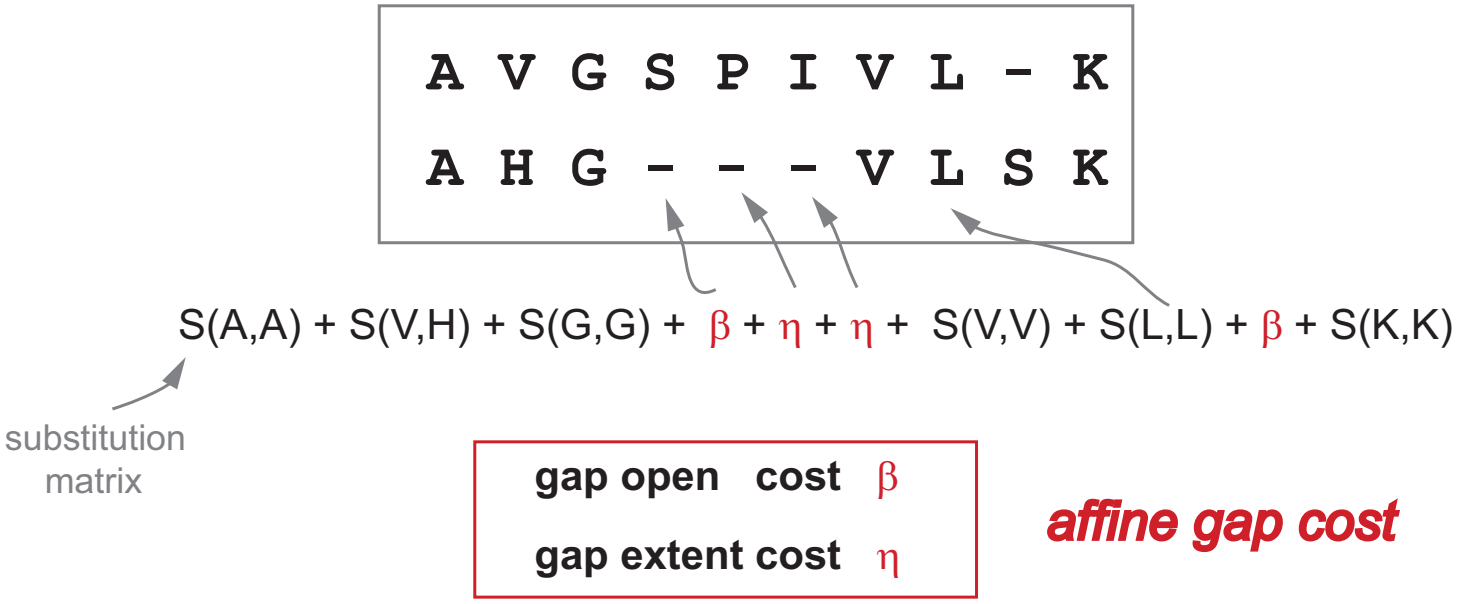
`Homology coverage`

Optimize for Evolutionary Distance

`Alignment of homologs`

Affine Gap cost

A way of dealing with variability

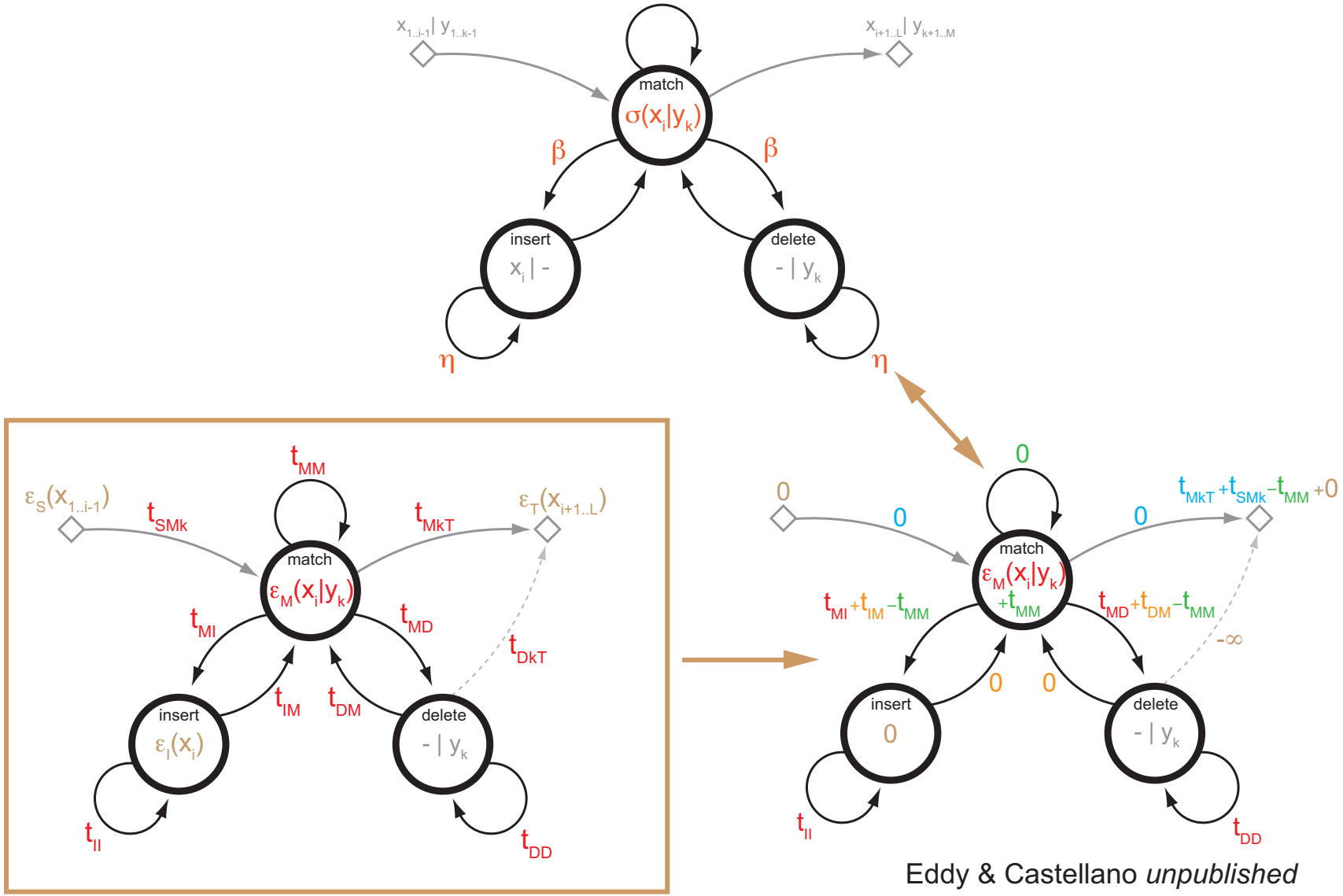


BLAST syncs (empirically) the choice of substitution matrix with that of the affine gap costs

substitution matrix	**BLOSUM62**
gap open	-11
gap extent	-1

HMMs formalize sw-like affine methods

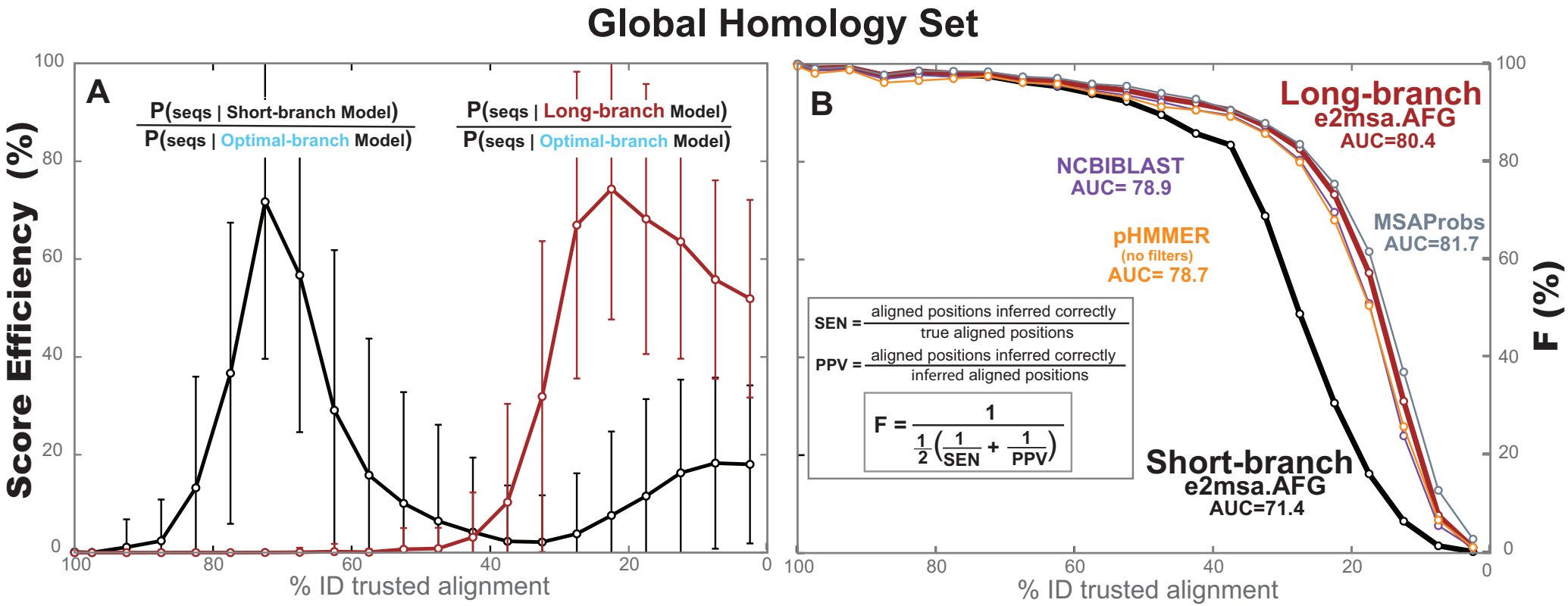
From Smith-Waterman to an HMM



A probabilistic evolutionary model provides time-dependent HMM transitions

Is it worth to parameterize pair and profile HMMs with an explicit evolutionary model?

Alignment Accuracy Benchmark



Evolution of residue substitutions

Assume

For very small times ε:

RATES

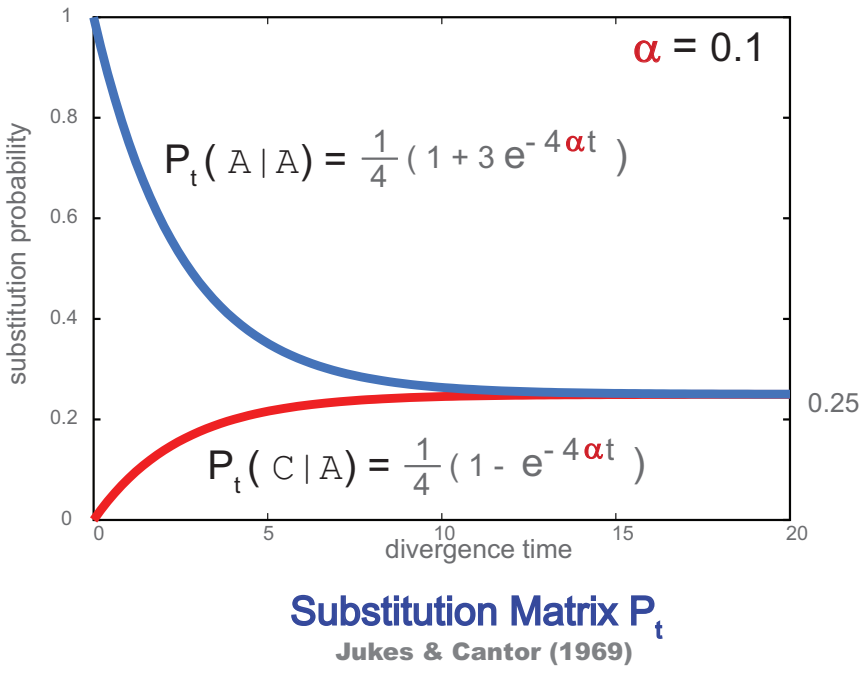
$\alpha \cdot \varepsilon$: A → C, A → G, A → T — residue changes

$-3\alpha \cdot \varepsilon$: A → A — residue is unchanged

propose and solve differential equations

Infer

For finite time *t:*

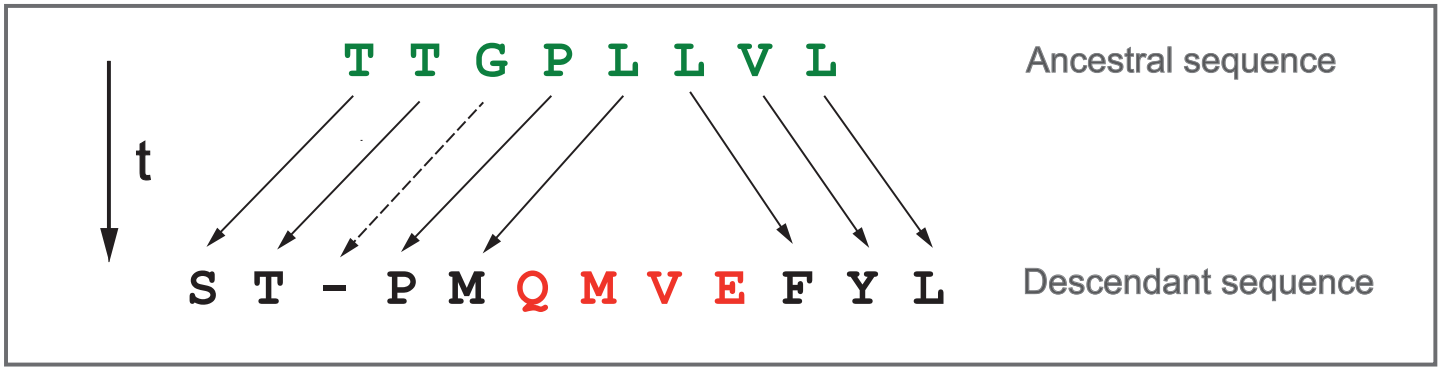


Substitution Matrix P_t

Jukes & Cantor (1969)

Evolution of Insertions

compatible with affine models



Substitutions

infinitesimal rate α

$P_t(\mathbf{S} \mid \mathbf{T})$

Insertions & Deletions

rate for **deleting** an ancestral residue μ_A

rate for **starting a new insert** with "n" residue $\lambda\,(1-s_I)\,s_I^{n-1}$

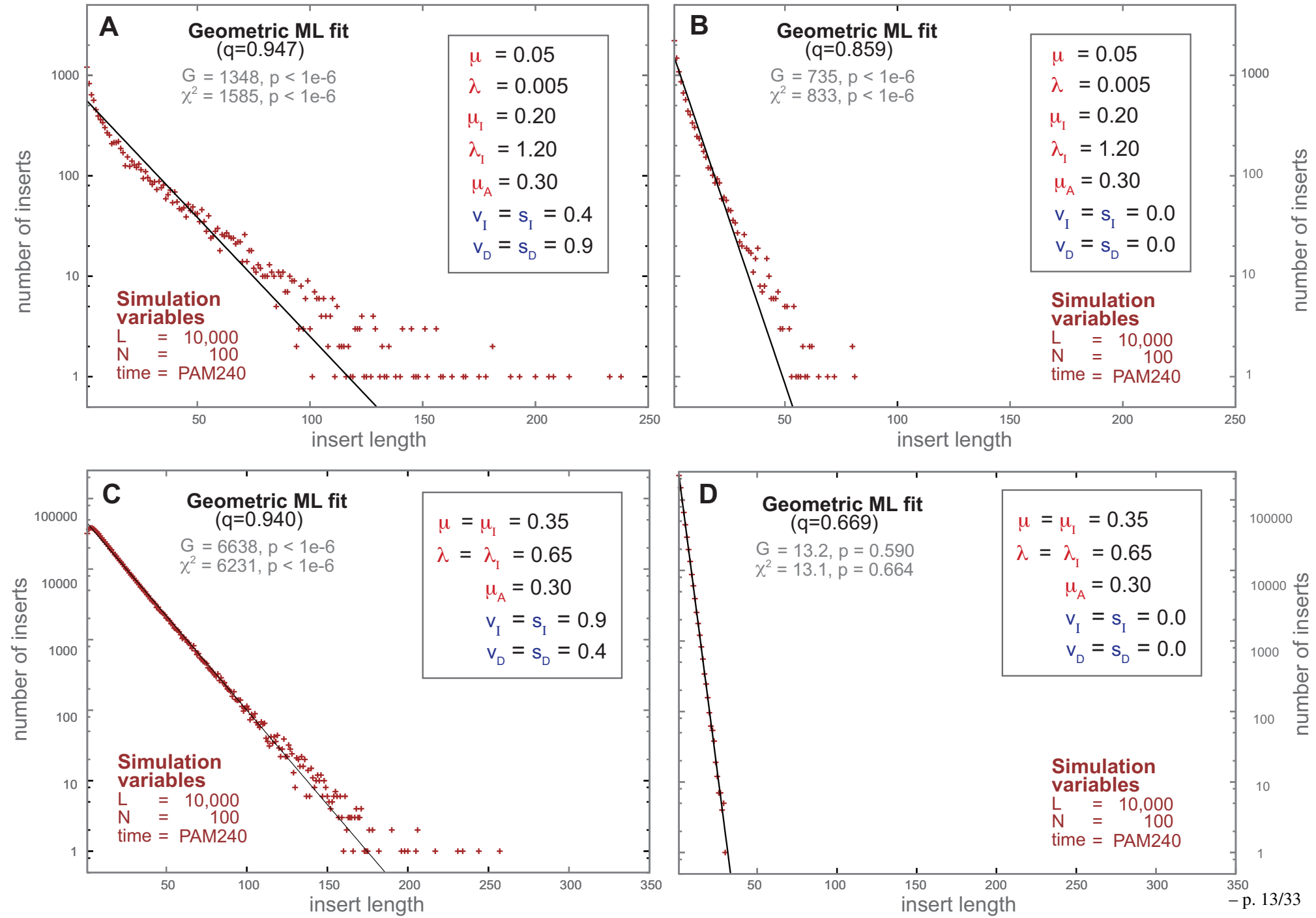
rate for **deleting a whole insert** with "n" residues $\mu\,(1-s_D)\,s_D^{n-1}$

rate for **adding to an insert** "x" residue $\lambda_I\,(1-\nu_I)\,\nu_I^{x-1}$

rate for **removing from an insert** "x" residues $\mu_I\,(1-\nu_D)\,\nu_D^{x-1}$

P_t (Descendant | Ancestral)

Not affine Models



Analytical closed-form solutions

AIF (fragment) Model

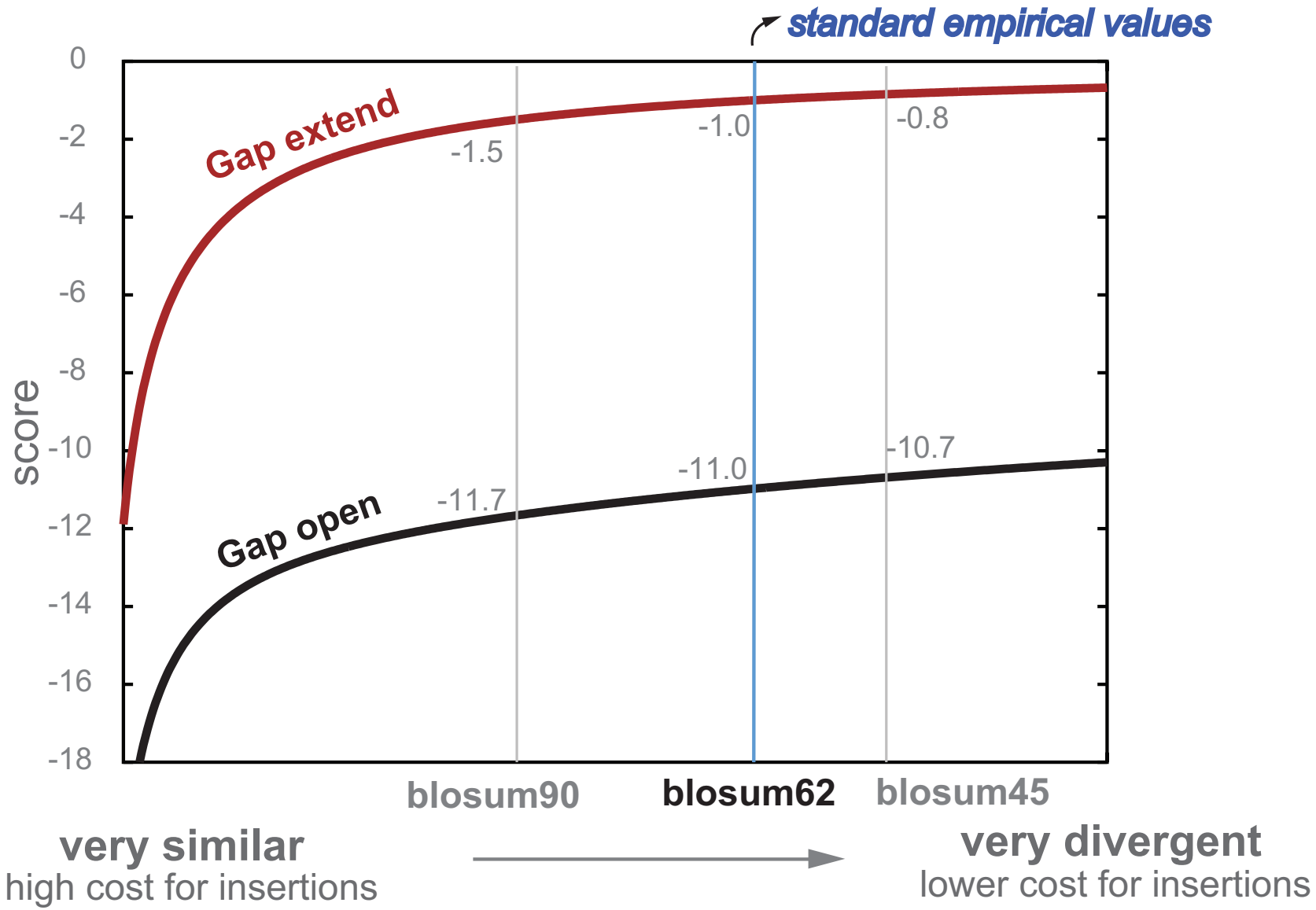
Gap opens: $\beta_t = \lambda_I \dfrac{1 - e^{(\lambda_I - \mu_I)\,t}}{\mu_I - \lambda_I e^{(\lambda_I - \mu_I)\,t}}$

Gap extends: $\eta_t = \dfrac{\lambda_I (1 - r) + \mu_I r - \lambda_I e^{(\lambda_I - \mu_I)\,t}}{\mu_I - \lambda_I e^{(\lambda_I - \mu_I)\,t}}$

Ancestral residue dies: $\gamma_t = 1 - e^{-(\mu_A)\,t}$

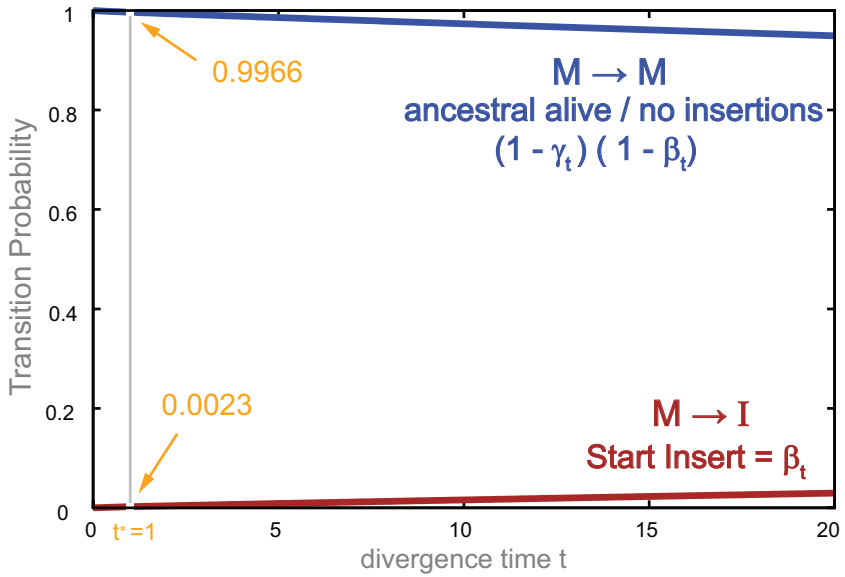
More realistic microscopic models result in non-affine macroscopic solutions

Evolved BLAST

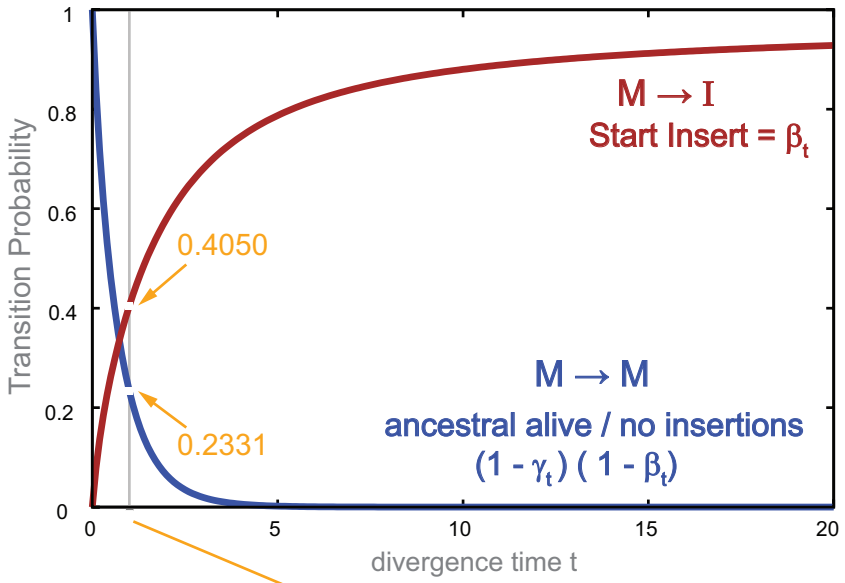


An evolved HMM

Position in a conserved region



Position at start of an insertion



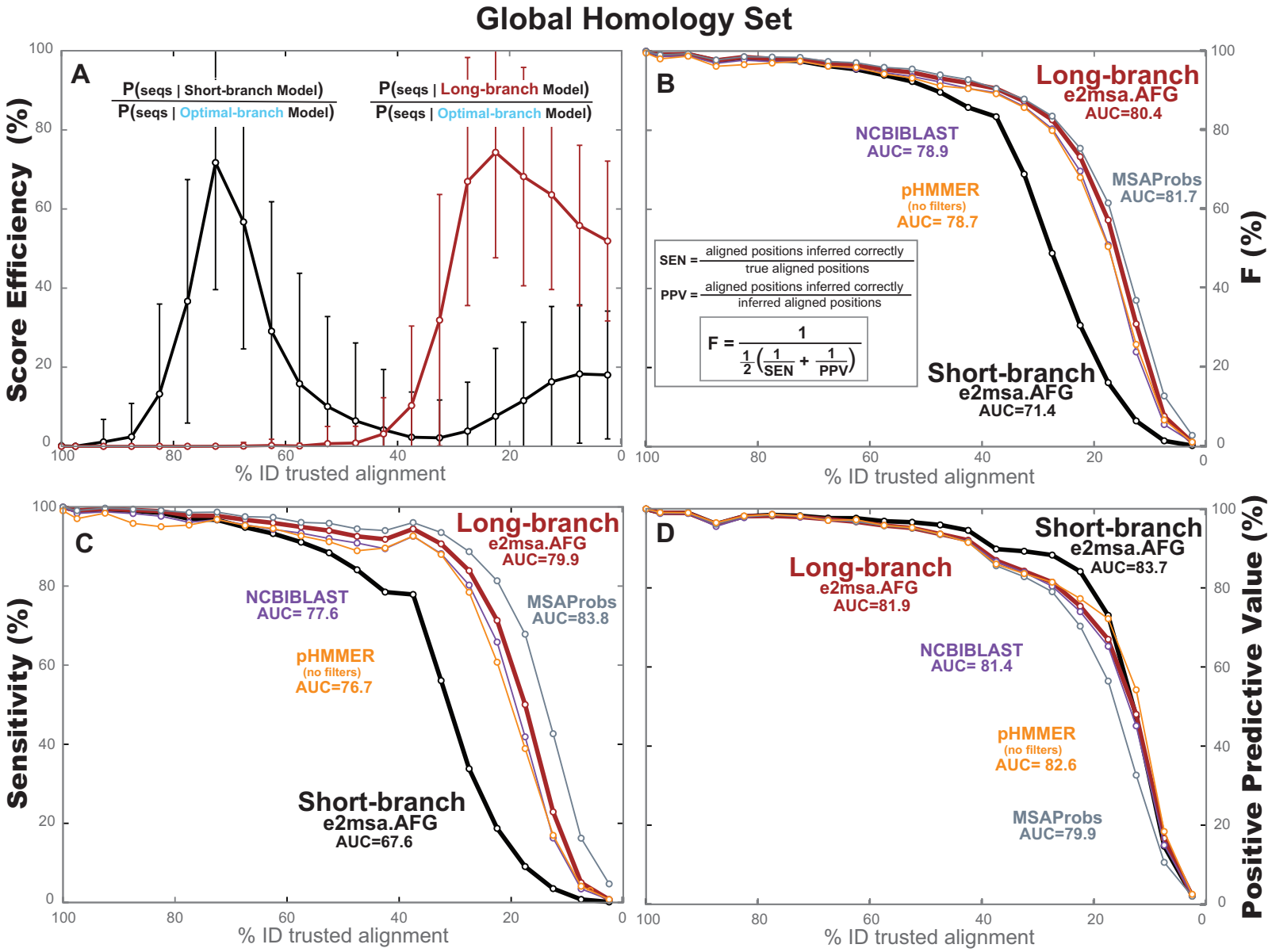
time at which parameters were trained from data

Affine Evolutionary Models

A Catalog

EVOLUTIONARY MODEL	Microscopic model: total # free parameters	Microscopic model: rates	Microscopic model: geometric parameters	Macroscopic model: # states minimal HMM	Macroscopic model: other properties
single-residue models					
AALI	6	$\lambda_I, \mu_I, \mu_A^{\{M,D,I\}}$	p	3	not reversible in general
LI	4	λ_I, μ_I, μ_A	p	1	not reversible in general
LR	2	$\lambda_I, \mu_A, (\mu_I = \lambda_I + \mu_A)$	$(p^{\mathrm{LR}} = \lambda_I/\mu_A)$	1	reversible
TKF91	2	λ, μ	$(p^{\mathrm{TKF}} = \lambda/\mu)$	2	reversible, ref. [**?**]
fragment models					
AFGX	9	$\lambda_I, \mu_I, \mu_A^{\{M,D,I\}}$	r_M, r_D, r_I, p	3	not reversible
AFG	7	λ_I, μ_I, μ_A	r_M, r_D, r_I, p	3	not reversible
AFGR	4	λ_I, μ_A	$r_M, r_{DI}, (p^{\mathrm{LR}})$	3	reversible
AFR	3	λ_I, μ_A	$r, (p^{\mathrm{LR}})$	3	reversible
TKF92	3	λ, μ	$r, (p^{\mathrm{TKF92}})$	3	reversible , ref. [**?**]
FID	2	λ	$r, (p = 1)$	3	reversible, ref. [**?**]
fragment affine model for profile HMMs					
AIFX	7	$\lambda_I, \mu_I, \mu_A^{\{M,D,I\}}$	r_I, p	3	not reversible
AIF	5	λ_I, μ_I, μ_A	r_I, p	3	not reversible
no-fragment affine model for profile HMMs (plan7 HMMER)					
AGAX	9	$\lambda_{\{M,D\}}, \mu_{\{M,D\}}, \mu_A^{\{M,D,I\}}$	s_I, p	3	not reversible
AGA	7	$\lambda_{\{M,D\}}, \mu_{\{M,D\}}, \mu_A$	s_I, p	3	not reversible

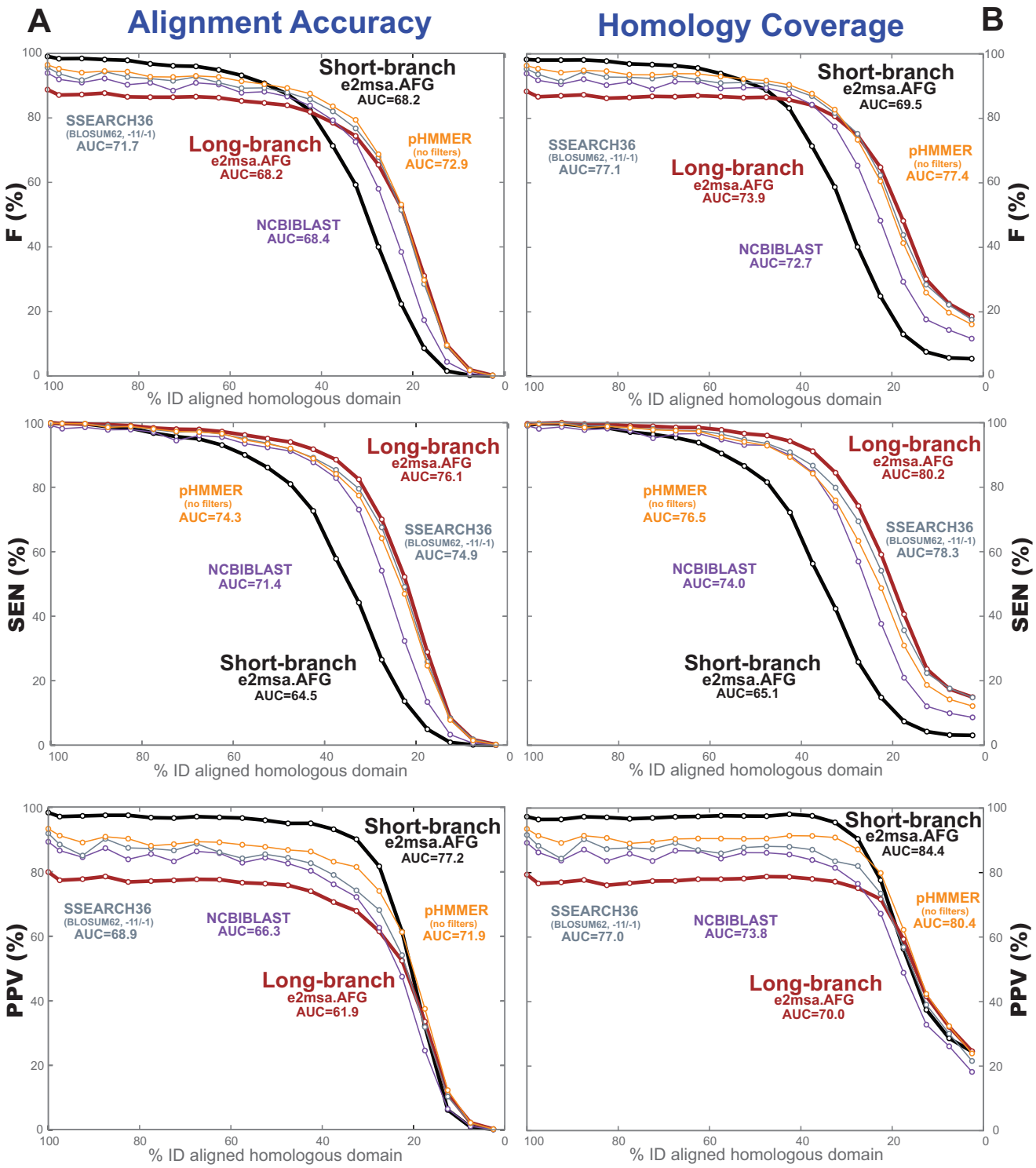
A **fixed *long-branch*** parameterization is sufficient to align ***global*** **homologies** of all degrees of conservation.


Global Homology Set
A
P(seqs | Short-branch Model) / P(seqs | Optimal-branch Model)
P(seqs | Long-branch Model) / P(seqs | Optimal-branch Model)
Score Efficiency (%)
% ID trusted alignment
B
Long-branch
e2msa.AFG
AUC=80.4
NCBIBLAST
AUC= 78.9
MSAProbs
AUC=81.7
pHMMER
(no filters)
AUC= 78.7
SEN = aligned positions inferred correctly / true aligned positions
PPV = aligned positions inferred correctly / inferred aligned positions
F = 1 / (1/2 (1/SEN + 1/PPV))
Short-branch
e2msa.AFG
AUC=71.4
F (%)
% ID trusted alignment
C
Long-branch
e2msa.AFG
AUC=79.9
NCBIBLAST
AUC= 77.6
MSAProbs
AUC=83.8
pHMMER
(no filters)
AUC=76.7
Short-branch
e2msa.AFG
AUC=67.6
Sensitivity (%)
% ID trusted alignment
D
Short-branch
e2msa.AFG
AUC=83.7
Long-branch
e2msa.AFG
AUC=81.9
NCBIBLAST
AUC= 81.4
pHMMER
(no filters)
AUC= 82.6
MSAProbs
AUC=79.9
Positive Predictive Value (%)
% ID trusted alignment

A **fixed *short-branch*** parameterization reduces non-homologous alignment overextension for ***high-identity local*** homologies.

50 amino acid homologies

Local Homology Set

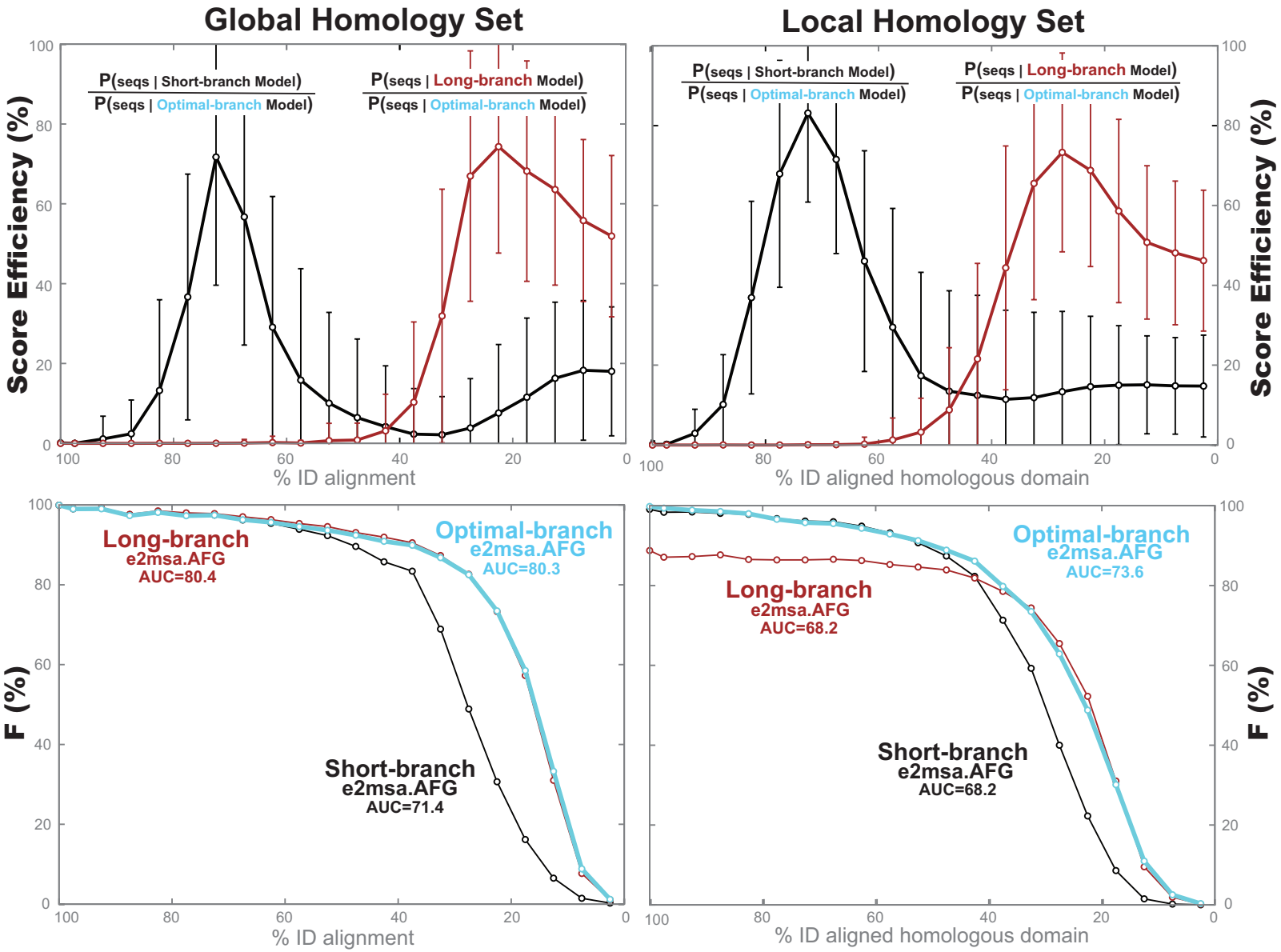


Optimal branch parameterization

A **variable** ***optimal-branch*** parameterization is best to align **local homologies** of any percentage identity.

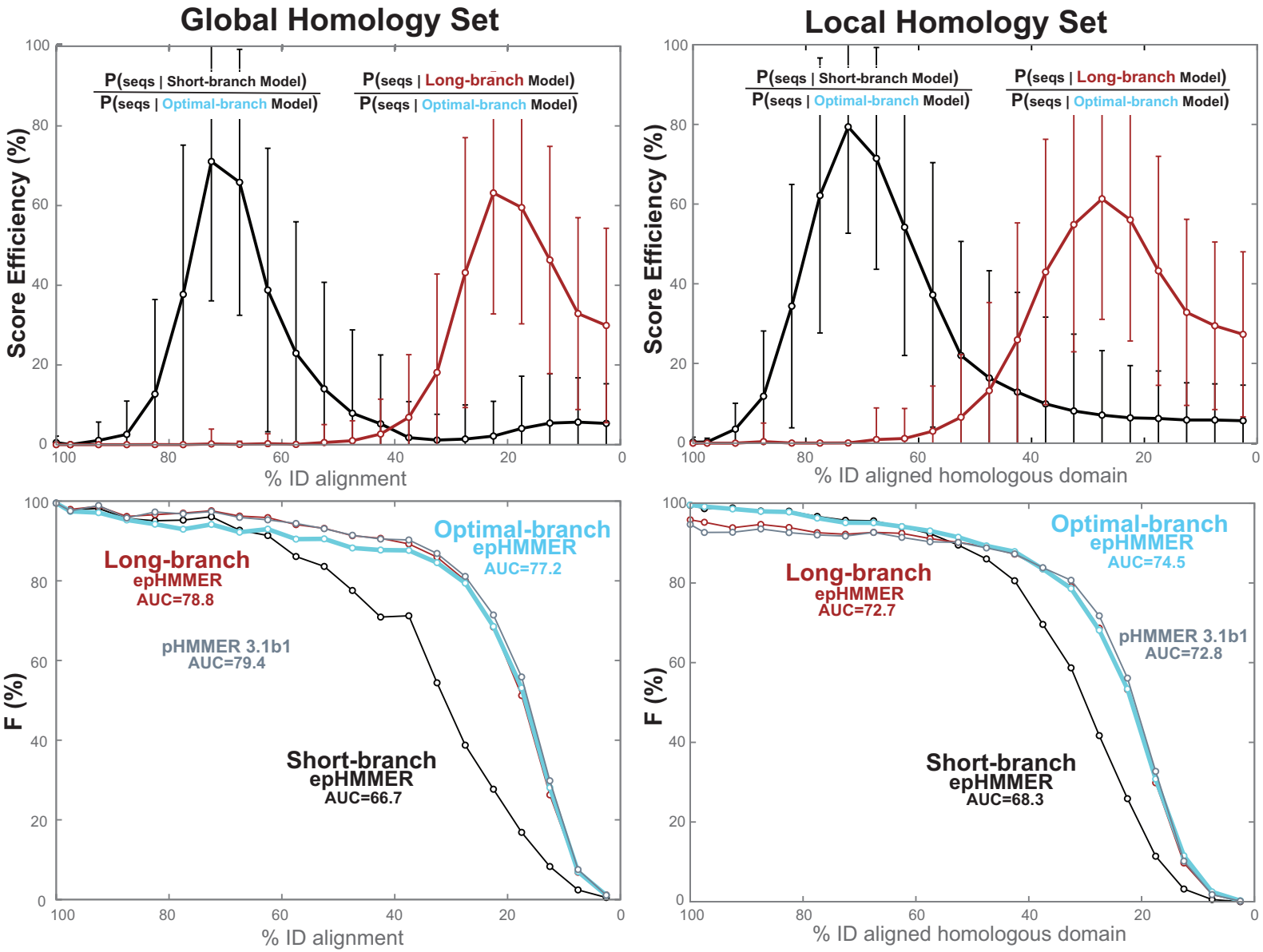
e2msa - pairHMM aligner

Alignment Accuracy- Evolutionary pair HMM (e2msa)



epHMMER

Alignment Acuracy - Evolutionary pHMMER (epHMMER)



Performance of different models

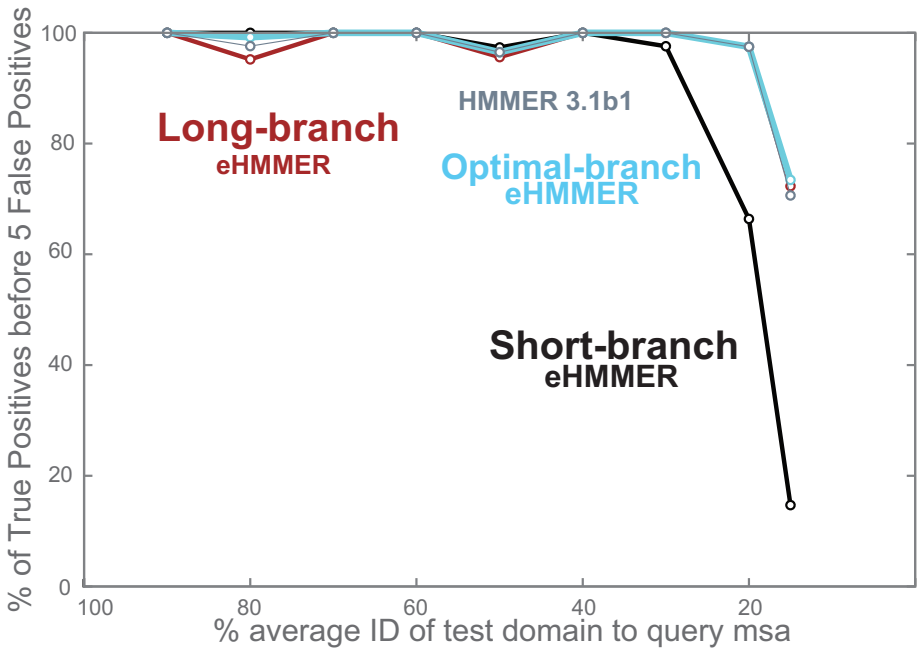
Method	Alignment Accuracy [AUC for F measure (%)]					
	Global Homology Set			**Local Homology Set**		
	PARAMETERIZATION			PARAMETERIZATION		
	SHORT	LONG	OPTIMAL	SHORT	LONG	OPTIMAL
`e2msa`.AFG	71.4	**80.4**	**80.3**	68.2	68.2	**73.6**
`e2msa`.AGA	71.3	**80.4**	**80.2**	68.1	67.3	**73.6**
`e2msa`.AIF	71.3	**80.4**	**80.2**	68.1	68.3	**73.3**
`e2msa`.TKF92	71.2	**80.0**	**79.9**	68.1	68.2	**73.4**
`e2msa`.LI	71.0	**78.7**	**78.6**	67.9	66.4	**72.7**
`e2msa`.TKF91	69.5	**75.4**	74.5	66.2	69.1	**70.7**
epHMMER (no filters)	66.7	**78.8**	77.2	68.3	72.7	**74.5**
pHMMER (no filters)	78.7			72.9		
SSEARCH (BLOSUM62, -11/-1)	80.0			71.7		
NCBIBLAST	78.9			68.4		
MSAProbs	81.7			36.1		
MUSCLE	80.8			33.5		

Evolutionary models with more parameters tend to perform better

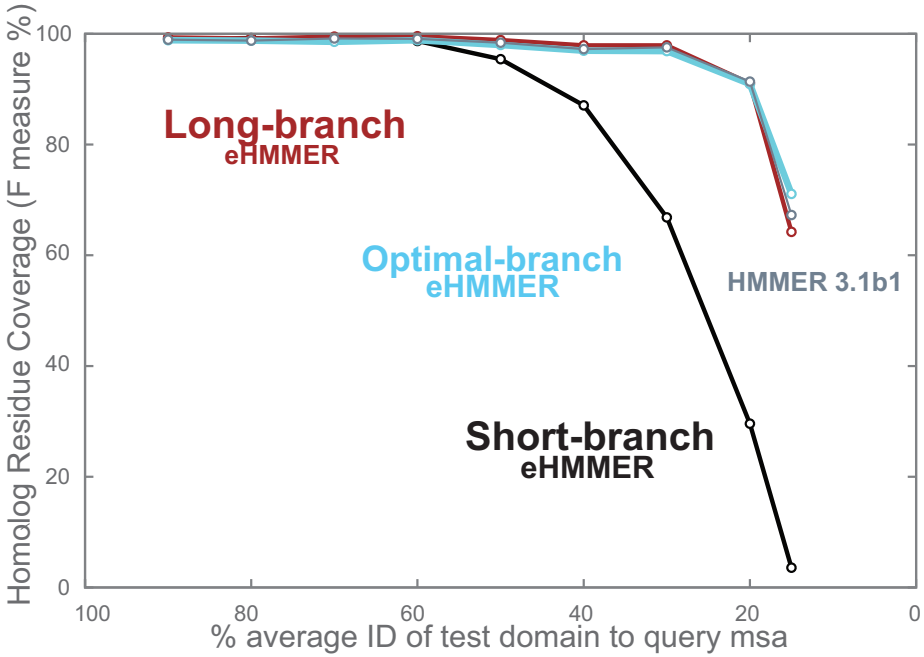
eHMMER

The *detection and coverage* of **embedded global homologies** is robust with just one **long-branch** parameterization

Embedded Global Homologies



Homology Detection

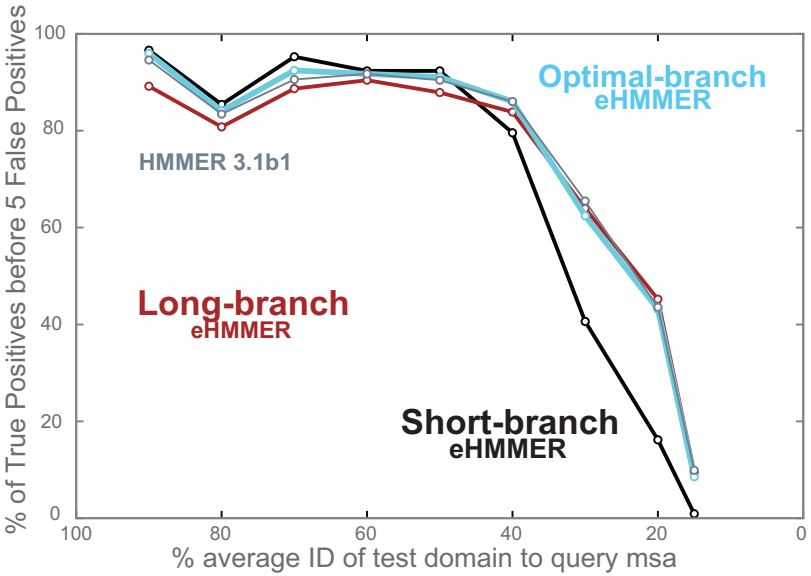


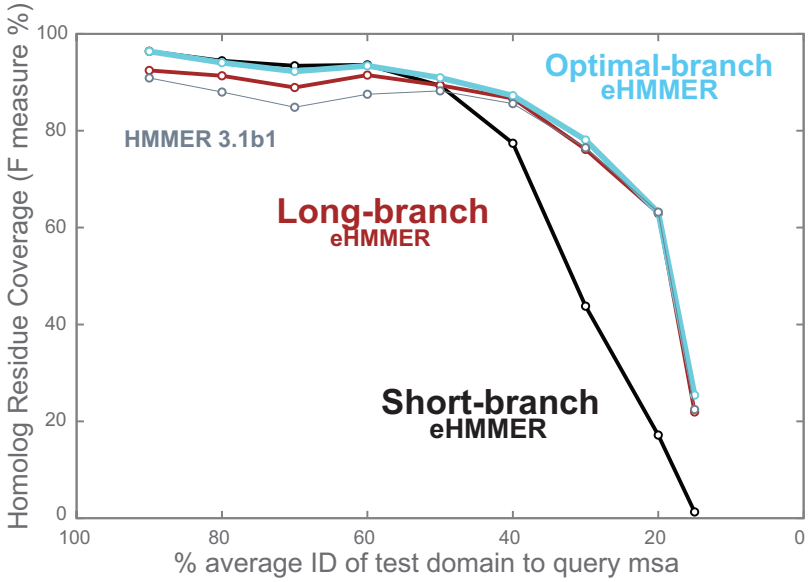
Homology Coverage

Short Local Homologies

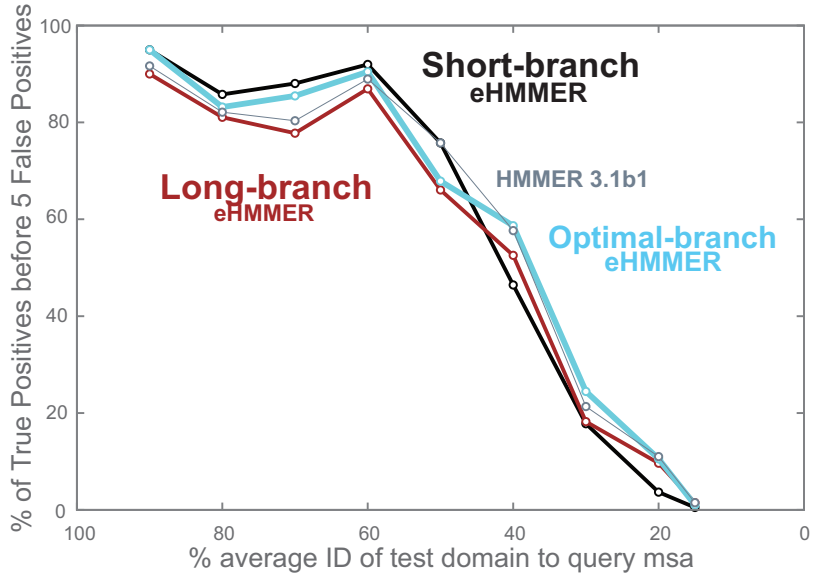
The *detection and coverage* of **embedded short local homologies** improves with a variable **optimal-branch** parameterization

Embedded 50 aa Local Homologies

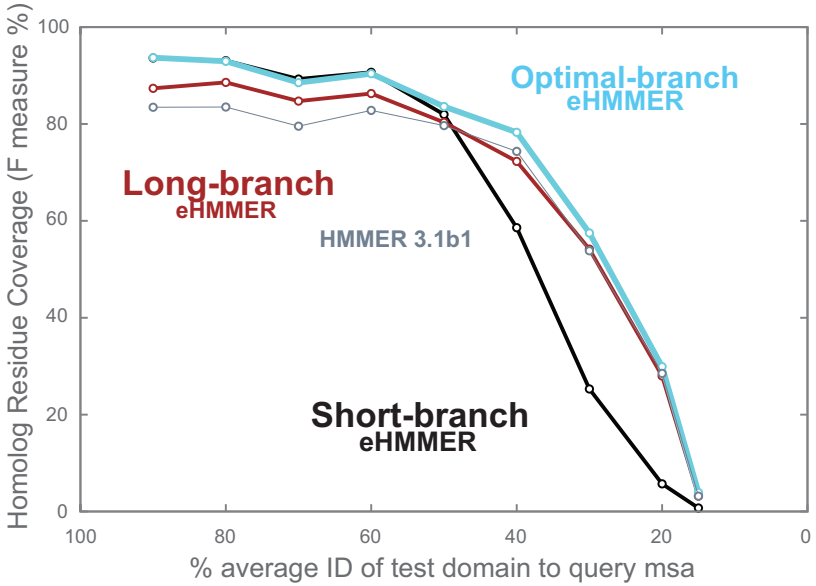




Embedded 30 aa Local Homologies



Homology Detection

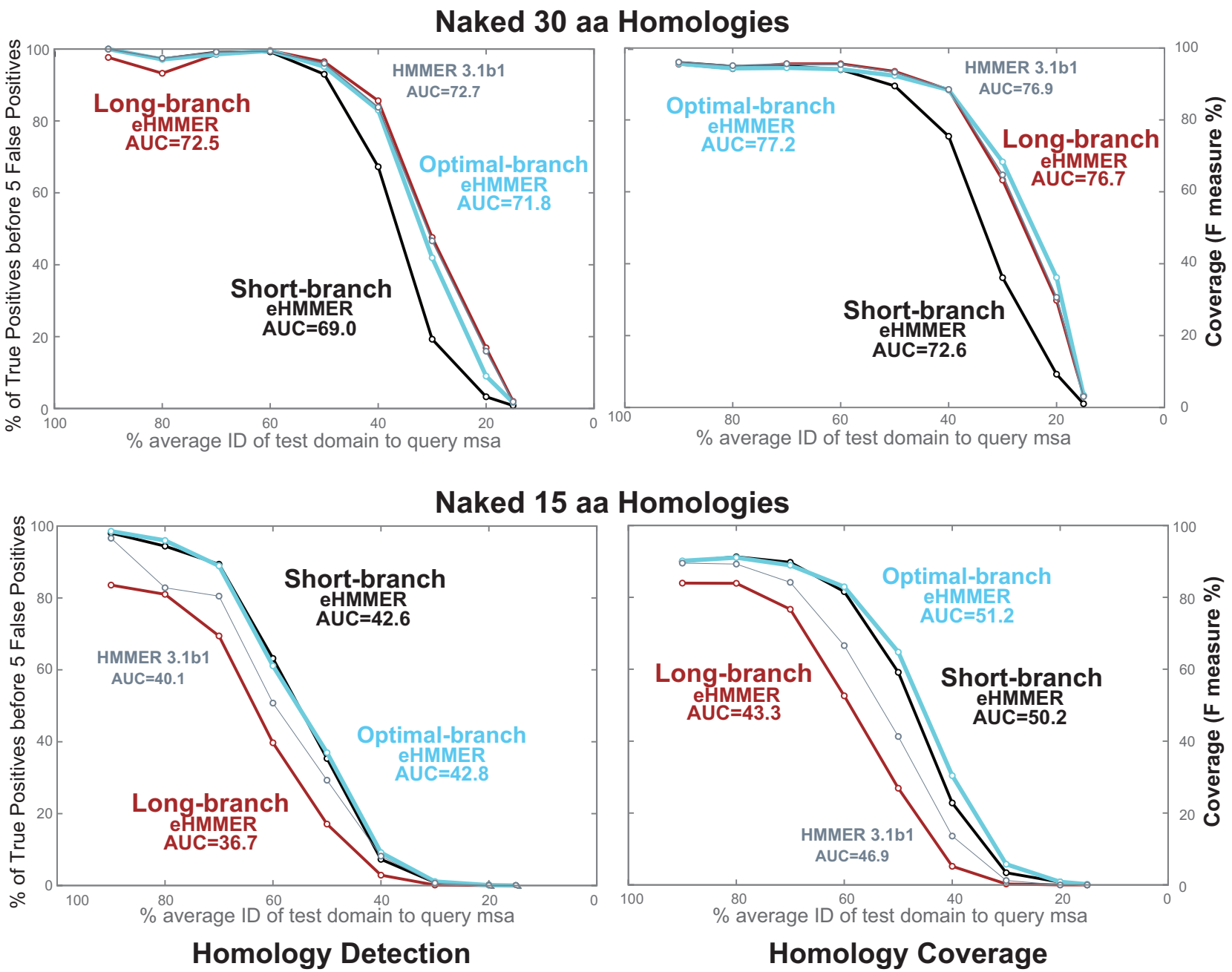


Homology Coverage

Fragments

The *detection* of **very short naked local homologies** improves with a **short-branch or optimal-branch** parameterization

Naked Fragments



Explicit evolutionary models??

It is nice to wind up and down a model without additional information

For Sensitivity —> Use a **long-branch** parameterization (12% id).
Except for metagenomics < 30 aa,
then use a **short-branch** parameterization (45% id).

For SEN/PPV —> Use a **optimal-branch** for short embbeded homologies.
For global embedded homologies still OK
using a **long-branch** parameterization.

Ancestral reconstruction