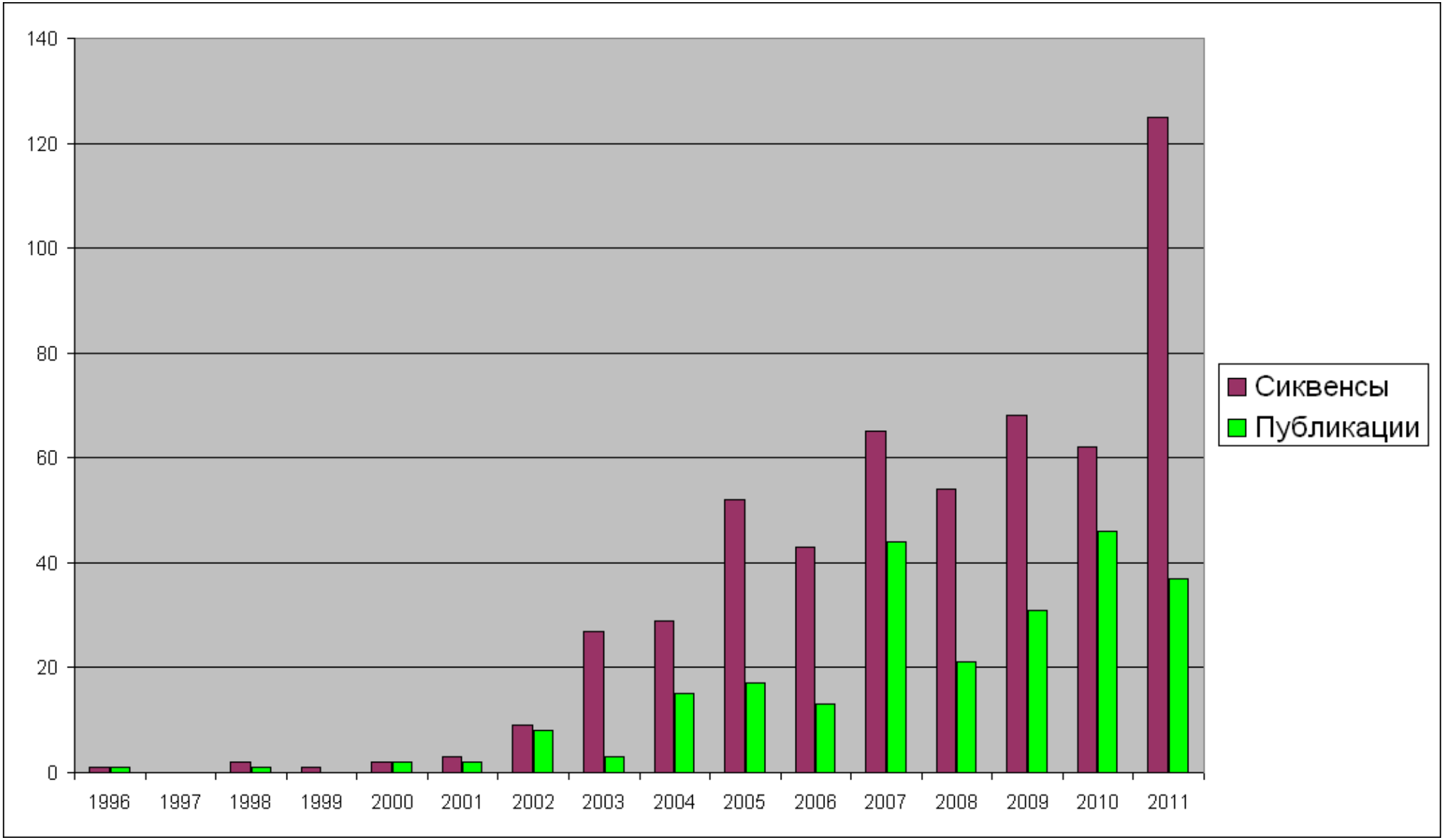
Аннотирование геномов

Кол-во секвенированных (NCBI) и опубликованных (wiki) геномов



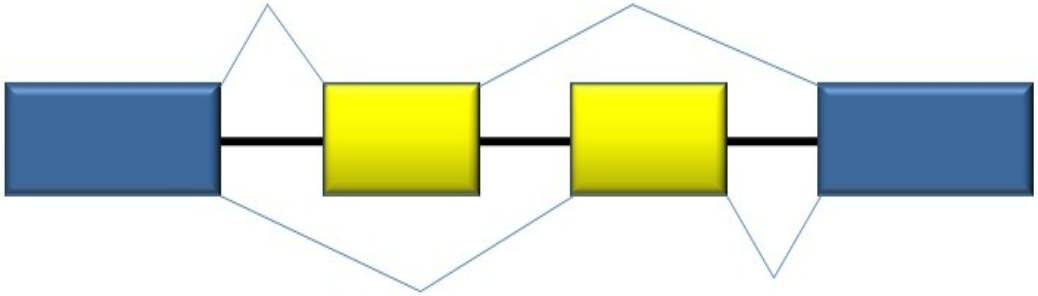
Геномы по группам (NCBI, wiki)

Группа	Секвенир. геномов	Аннотир. за 2011 г.
Животные	438	12
Грибы	413	4
Простейшие	210	2
Растения	154	7
Всего	1215	25

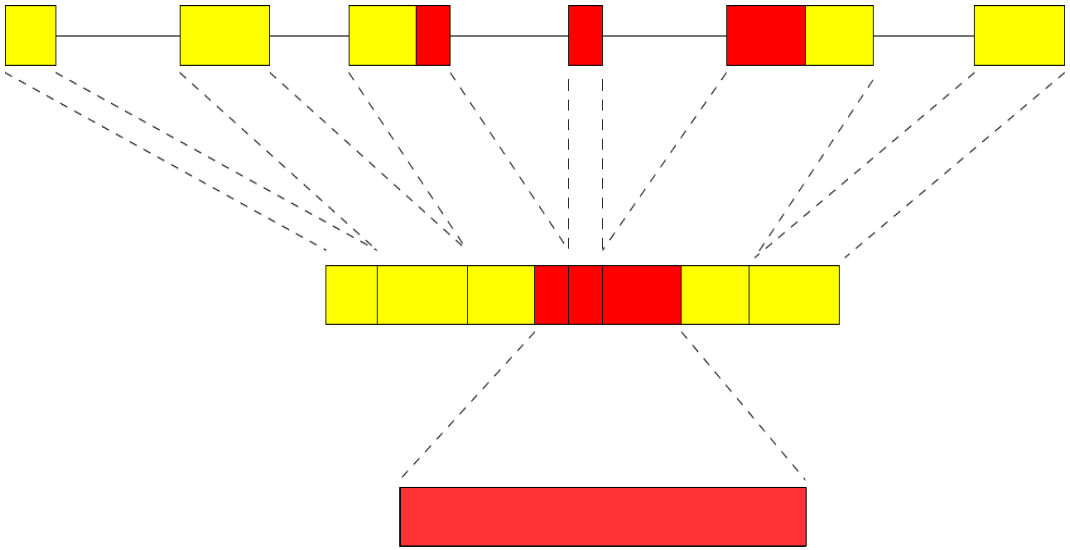
Процесс аннотации генома

- Поиск повторов и транспозонов
- **Поиск генов (экзонов и транскриптов)**
  - Предсказание Ab initio
  - Предсказание с использованием гомологии/экспрессии
  - Соединение нескольких предсказаний в одно
- **Поиск стренда и транслируемой области**
- Поиск гомологов/ортологов, функциональных доменов, классификации по GO, KEGG, ...
- Поиск некодирующих РНК

Задача



- Поиск экзонов и объединение их в наборы (транскрипты)



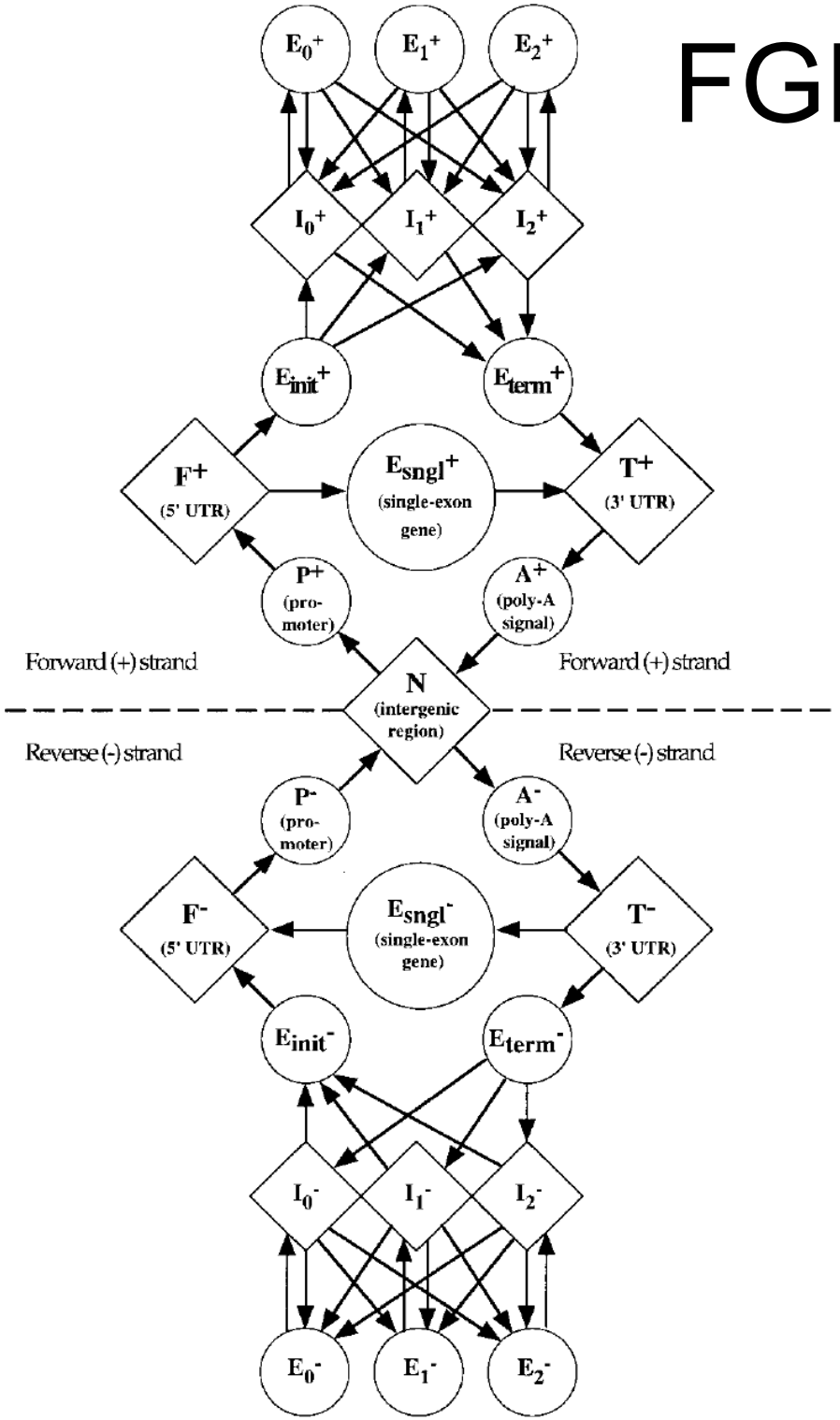
- Поиск транслируемой области транскрипта

Group	Genome	Ab initio	Transcriptome	Evidence based	Web pipeline
Vertebrate	Meleagris gallopavo	Twinscan, GlimmerHMM	-	JIGSAW	Ensembl
Vertebrate	Gorilla gorilla	-	-	-	Ensembl
Vertebrate	Pongo pygmaeus	-	-	Exonerate	Ensembl
Worm	Ascaris suum	Augustus, GlimmerHMM, SNAP, BestORF	-	genBlastA, Genewise, Cufflinks, Glean	-
Worm	Clonorchis sinensis	Augustus, Genscan	-	genBlastA, GeneWise, GMAP, EvidenceModeler	-
Worm	Schistosoma haematobium	Augustus, GlimmerHMM, SNAP	SOAPdenovo	Glean	-
Spider	Tetranychus urticae	SpliceMachine, EuGene	-	EuGene	EuGene
Insect	Acromyrmex echinatior	Augustus, SNAP	Cufflinks	GLEAN, Genewis	Apollo
Insect	Atta cephalotes	Augustus, Snap, GeneMark	-	Exonerate, MAKER	Apollo
Insect	Linepithema humile	Augustus, Snap, GeneMark	-	MAKER	Apollo
Insect	Pogonomyrmex barbatus	Augustus, Snap, GeneMark	-	MAKER	Apollo
Insect	Solenopsis invicta	Augustus, Snap, GeneMark	-	MAKER, Exonerate, Genewise	Apollo
Plant	Brassica rapa	Genscan, Augustus	PASA	GLEAN	-
Plant	Cannabis sativa	-	AbySS	Inhouse	-
Plant	Fragaria vesca	GeneMark-ES+	TGICL	GeneMark-ES+	-
Plant	Medicago truncatula	SpliceMachine, FGENESH	PASA, Cufflinks	Eu'Gene	-
Plant	Phoenix dactylifera	Fgenesh++	Fgenesh++	Fgenesh++	-
Plant	Selaginella moellendorffii	Fgenesh+, Genewise	JGI pipeline	JGI pipeline	JGI pipeline
Plant	Thellungiella parvula	FGENESH++, Genescan	FGENESH++	FGENESH++	-
Fungi	Leptosphaeria maculans	Eugene_IMM, SpliceMachine, Fgenesh	-	URGI, GenomeThreader	-
Fungi	Mortierella alpina	Augustus, GlimmerHMM, SNAP, GeneMark	-	EvidenceModeler	-
Fungi	Puccinia graminis	Fgenesh, Genewise, estExt, EuGene, GeneID, Augustus	-	Fgenesh, Genewise, estExt, EuGene	JGI pipeline
Fungi	Serpula lacrymans	FGENESH, GeneMark	-	FGENESH+, Genewise	JGI pipeline
Protist	Albugo laibachii	GeneMark, RACE	-	Augustus	-
Protist	Aureococcus anophaefferens	FGENESH	-	FGENESH+, Genewise	-

Программы, чаще всего используемые в 2011 году

Программа	# использований
Fgenesh	15
Augustus	12
Genewise	9
GeneMark	7
Snap	7
EuGene	6
Glean	4
GlimmerHMM	4
Maker	4
Cufflinks	3
Exonerate	3
Genscan	3
SpliceMachine	3

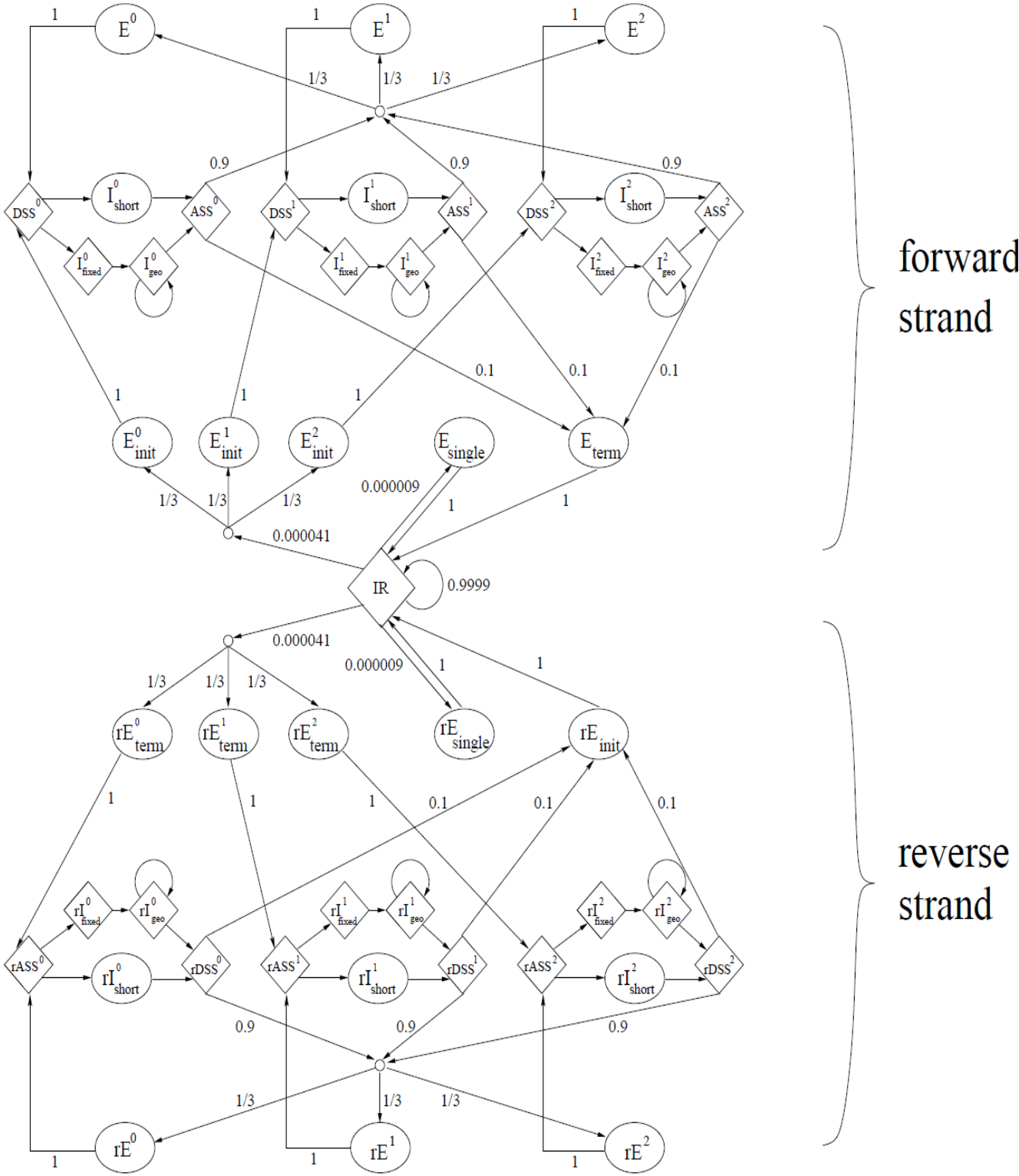
FGENESH+ (GENSCAN)

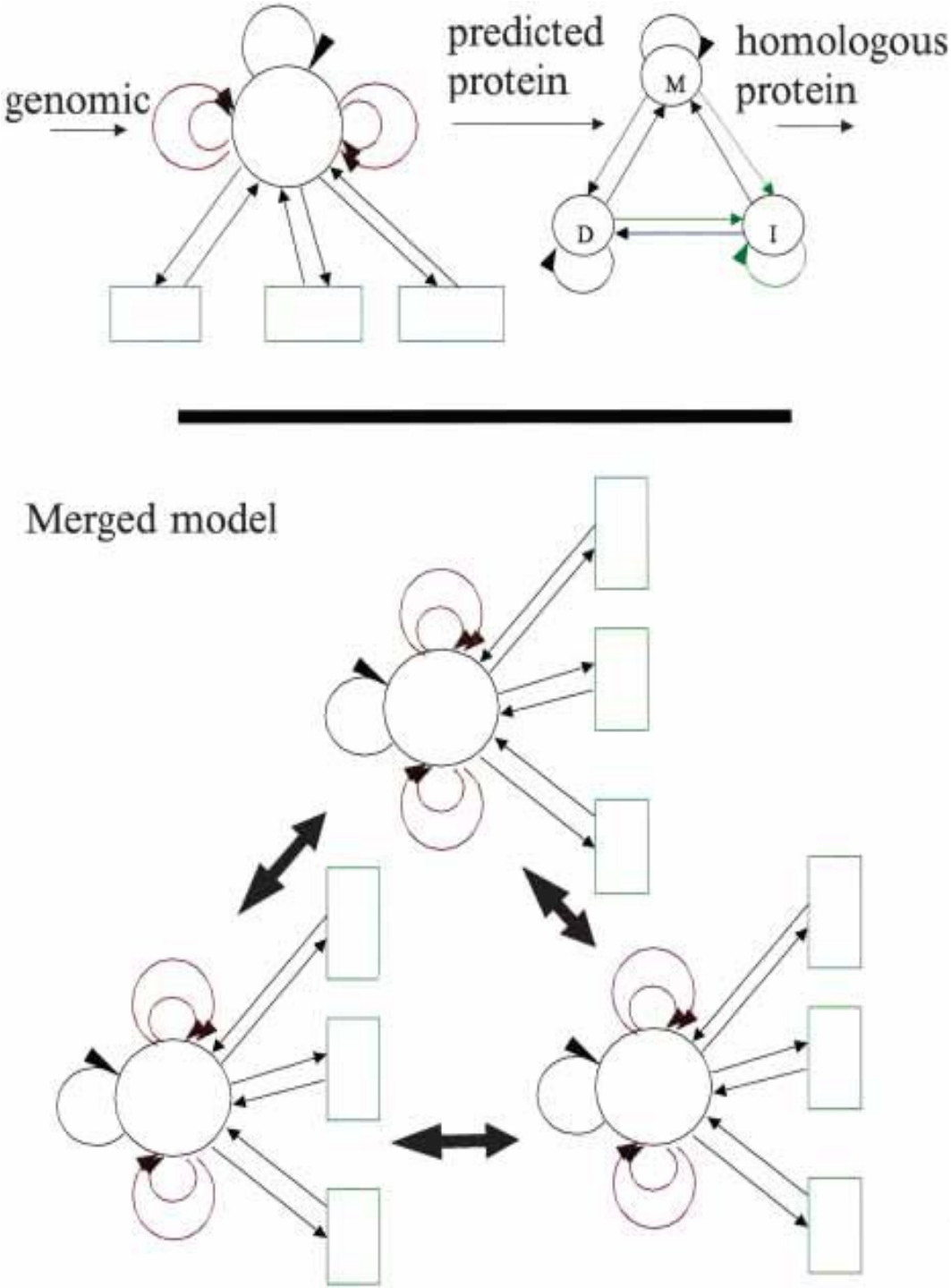


- Веса старта гена и сайтов сплайсинга зависят от днк вокруг сайта
- (+) - добавочная информация о гомологии с белками (SW для каждого потенциальн. экзона + штраф за делеции в белке между экзонами)

AUGUSTUS

- Модель интрона: не геометрическое распределение длин (Generalized HMM)
- Возможность учитывать (в весах наблюдений) информацию о выравнивании с белками и RnaSeq ридами (454, Illumina, Solid,...)





GeneWise

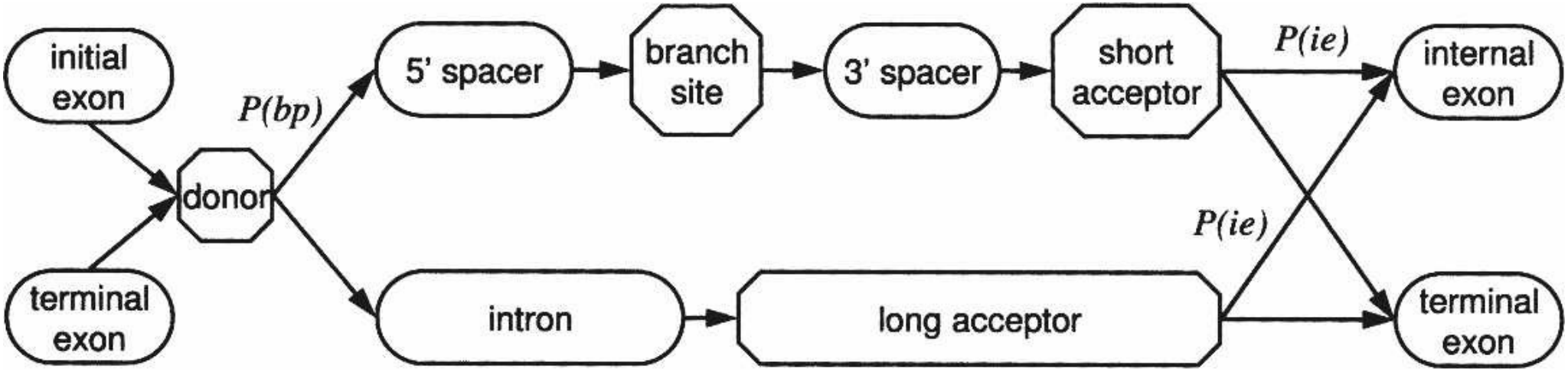
- 2 парные HMM:

  (1) генерирует ДНК + белок (структура гена),

  (2) генерирует выравнивание с другим белком

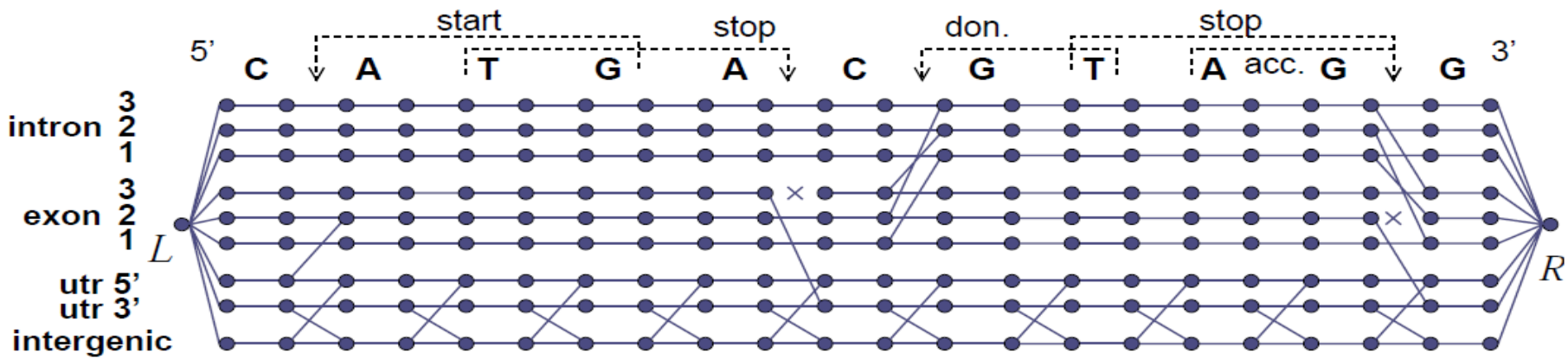
- Метод GeneWise используется в Ensembl pipeline

GeneMark



- Только ab initio
- Алгоритм не требует тренировочных данных
- Еще более сложная модель интрона
- Gibbs sampling (MCMC) «branch point» сайта
- Viterbi обучение

EuGene



- Динамическое программирование на графе из экзонных, интронных и меж-генной ветвей
- Веса сплайсинговых переходов оцениваются через внешние программы, веса экзонов оцениваются через IMM (ab initio) и увеличиваются, если есть наблюдения гомологии с белковым банком или Rna-Seq
- Нет стадии обучения

GlimmerHMM

- Используется Interpolated Markov Model (смесь марковских цепей разных порядков)
- Для предсказания сайтов сплайсинга используются марковские модели второго порядка и Maximal Dependence Decomposition (метод класса Decision Tree)
- MDD основан на χ^2-статистике взаимосвязи нуклеотидов в одних позициях сайта и индикаторов консенсусных букв в других позициях

GLEAN

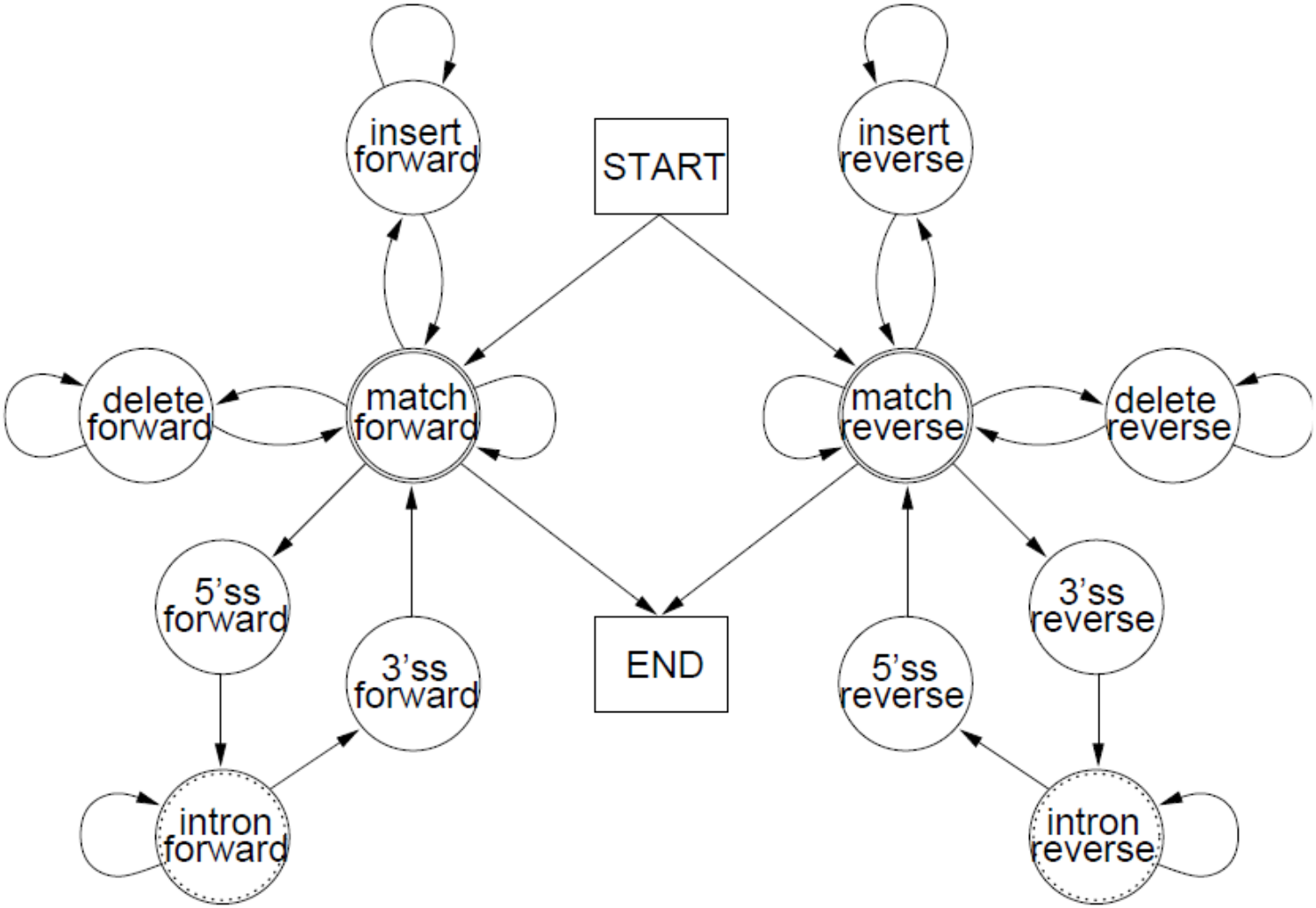
- Соединение нескольких генных разметок (это могут быть предсказания ab initio, по гомологии с белковыми банками, по профилям экспрессии из данных Rna-Seq)
- Latent Class Analysis: для каждого гена оцениваются апостериорные вероятности сайтов старт/стоп транс[крип/ля]ции и сайтов сплайсинга; HMM находит разметку с наибольшим правдоподобием
- Минус: недопредсказание (повисание генов на краях контигов)

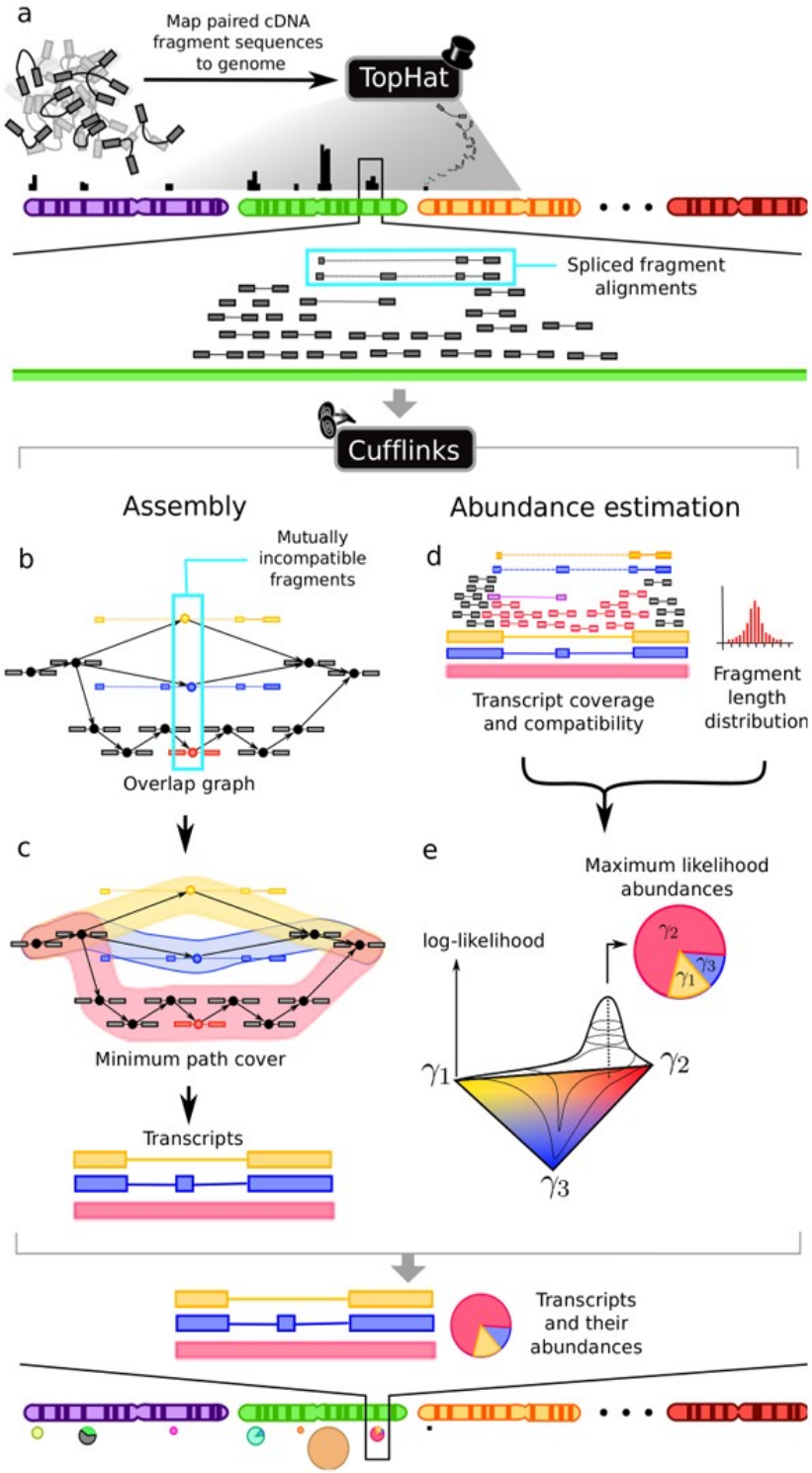
SpliceMachine

- Используется линейные SVM
- При обучении и предсказании учитываются следующие переменные: (1) частоты нуклеотидов, 2-меров, 3-меров в конкретных позициях -6,+7 вокруг GT/AG, (2) частоты 4-, 5-, 6-меров в более широкой области без привязки к позиции, (3) частоты кодонов слева и справа от ядра сайта в 3 рамках
- В многомерном пространстве всех этих переменных строится гиперплоскость

Exonerate

- Выравнивает РНК/Rna-Seq/EST/гены/белки на ДНК

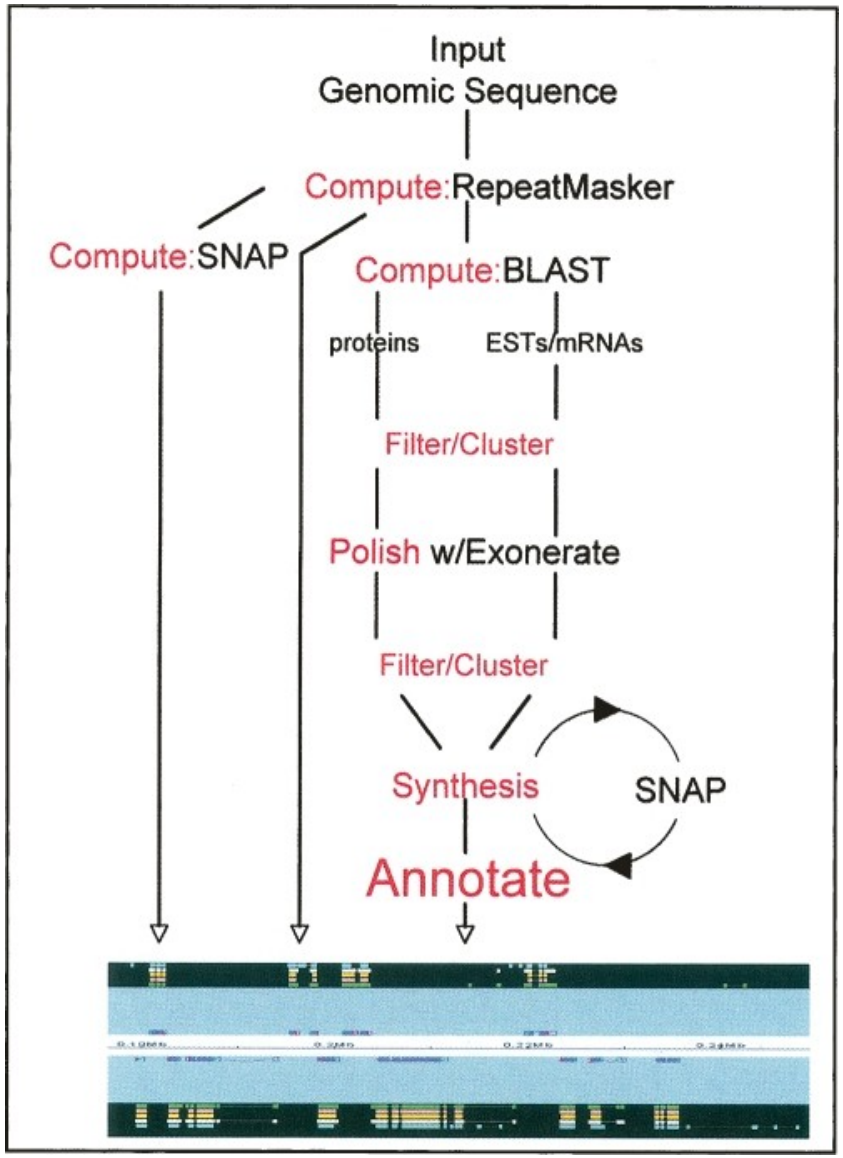




TopHat + Cufflinks

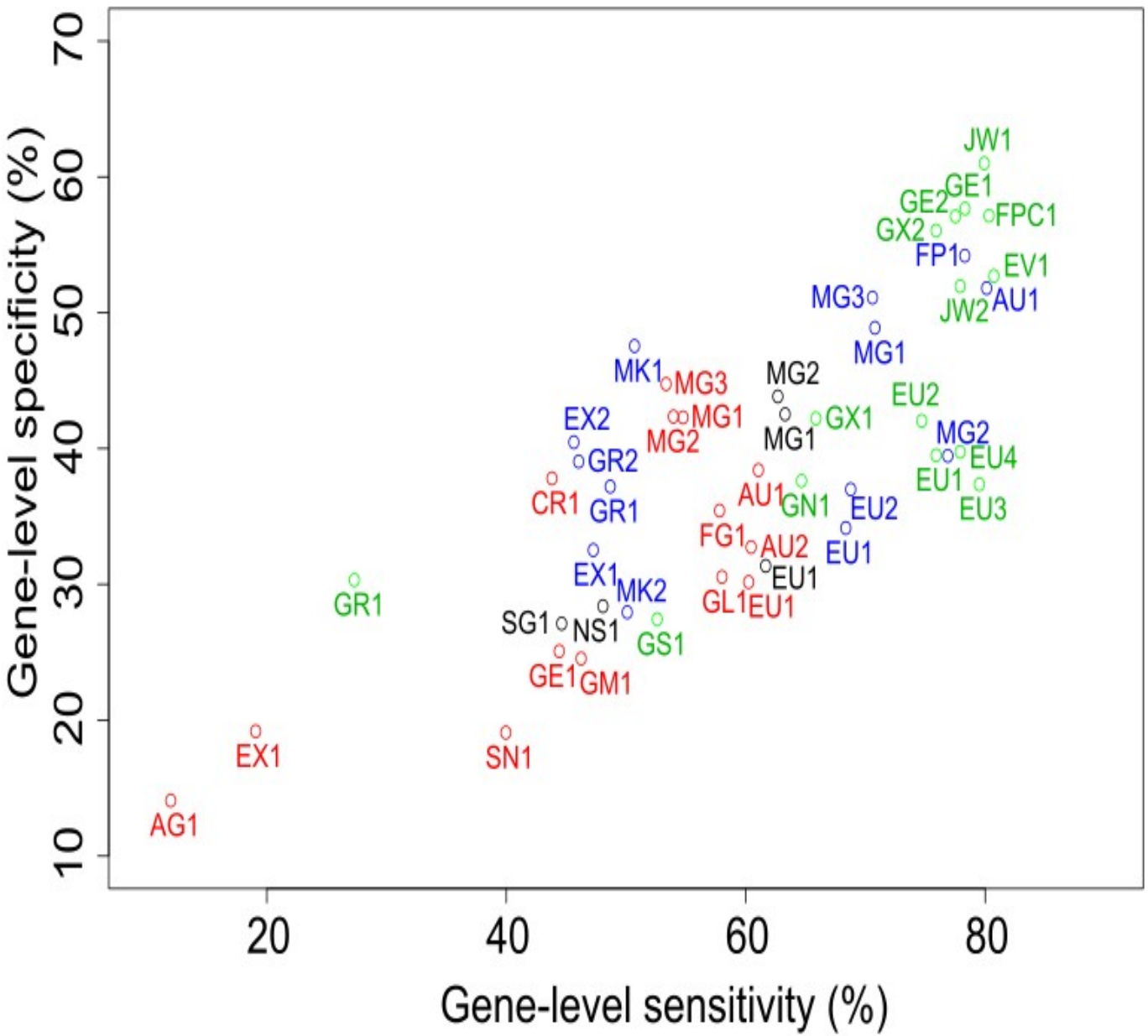
- TopHat выравнивает риды (которые **не выравниваются** Bowtie) методом seed-and-extend, идентичность должна быть близка к 100%, но зато скорость работы большая
- Cufflinks собирает риды в транскрипты (максимальные пути во взвешенном двудольном графе); далее частоты каждого из альтернативных транскриптов вычисляются максимизацией **правдоподобия всех ридов** данной группы транскриптов

MAKER



- Позиционируют себя как для старта проекта и для геномов новых таксономических групп
- В 2011 году появился MAKER2, использует SNAP, AUGUSTUS и GeneMark
- Набор перловых скриптов

Оценка качества (nGASP, C. elegans)



JW = JIGSAW

GE = GENEID

EV = EVIGAN

GX = GENOMIX

FP/FG = FGENESH

AU = AUGUSTUS

MG = MGENE

EU = EUGENE

GN = GLEAN

MK = MAKER (…)

GL = GlimmerHMM

GM = GENEMARK

SN = SNAP