Eukaryotic Gene Structure Prediction Using Duration HMM

Duration HMM을 이용한 진핵생물 유전자 구조 예측

Hongseok Tae and Kiejung Park*

Information Technology Institute, SmallSoft Co., Ltd., Daejeon, Korea

***To whom correspondence should be addressed. E-mail: kjpark@smallsoft.co.kr**

초록

주어진 염기서열에서 유전자 영역을 예측하는 유전자 구조 예측은 유전체 프로젝트의 중요한 과정 중 하나이며 유전체 프로젝트 전체에 큰 영향을 준다. 진핵생물의 유전체가 원핵생물의 유전체에 비해 더 복잡한 구조를 가지기 때문에 진핵생물의 유전자 구조 예측 모델 역시 원핵생물에 비해 다양한 모델이 제안되었다. 본 연구팀은 duration hidden markov model을 기본형태로 하여 EGSP(Eukaryotic Gene Structure Prediction)프로그램을 개발하였다. 현재 개발된 진핵생물의 유전자 구조 예측 알고리즘 중에서 GenScan이 가장 정교한 것으로 보고 되고 있는데, EGSP의 결과분석을 위해 GenScan과 함께 GeneID, Morgan의 예측결과를 여러 가지 기준에서 비교하였다. EGSP는 정교한 예측모델을 가지고 있음에도 각 구성모듈에 대한 파라메터의 정교함에서 부족한 면이 나타나므로, 모델의 개선과 각 모듈의 조율을 통해 더욱 개선된 결과를 가지게 될 것이다.

서론

유전체의 전반적인 구조와 기능을 밝히고자 하는 유전체 프로젝트가 시작된 이후 많은 생명체의 유전체에 대한 연구가 진행되고 있으며 그 결과가 데이터베이스로 저장되고 있다. 유전체 프로젝트의 첫 번째 단계라고 할 수 있는 유전체 염기서열 분석의 비율이 증가하면서 유전체 내의 정확한 유전자 위치를 알아내기 위해 많은 유전자 구조 예측 모델들이 개발되었다. 생물체의 유전체에 존재하는 유전자의 위치를 정확하게 밝혀내는 것은 유전자간의 연관성, 그 유전자로부터 얻어지는 단백질간의 연관성, 그리고 나아가서는 비슷한 유전자들을 가지는 생물종간의 연관성을 밝히는 전체 과정에서 가장 핵심적인 단계로서 매우 중요한 의미를 가진다.

1980년대 초에 Shepherd[1], Fickett [2], 그리고 Staden & McLachlan[3]에 의한 유전자 구조 예측의 초기연구에서는 아미노산 분포와 codon usage의 경향을 통계적으로 측정해서 genome sequence에 존재하는 단백질의 coding region을 밝혀내고자 하였다. 그 후 k-tuple frequencies [4], autocorrelation[5], Fourier spectra [6], purine/pyrimidine periodicity[7], 그리고 local compositional complexity/

entropy[8]등 coding region과 non-coding region에서의 차별적인 구성을 기지는 특성들이 알려지면서, 이러한 구성들의 차이를 이용하여 유전체에 존재하는 coding region의 정확한 위치를 밝혀내고자 하는 시도가 이루어졌고 이와 더불어 유전자 구조 예측 프로그램들이 등장하기 시작했다. 그 중에서 Fickett[2]의 모델에 근거한 TestCode와, neural network 접근방식으로 여러 가지 구성에 대한 통계적 수치를 적용해 염기서열 단편을 coding region과 non-coding region으로 구분한 GRAIL[9]이 가장 널리 사용되었다.

이전까지 만들어진 유전자 구조 예측 모델들은 DNA 한쪽 가닥만을 분석하도록 만들어졌지만, GenMark[10]는 DNA 두 가닥을 동시에 분석하여 한쪽 가닥의 coding region에 의해서 다른 가닥에서도 그 위치에서 non-coding region임에도 불구하고 coding region처럼 인식되는 'shadow' coding region 문제를 해결하고자 하였다. GenMark는 non-coding region에서는 homogeneous 5th-markov chain, coding region에서는 codon의 위치특이적인 non-homogeneous 5th-markov chain을 DNA 양쪽 가닥에 모두 구성하고, 각 markov chain의 상대적인 score에 따라 coding region을 찾아낸다. GenMark 이후 원핵생물 유전체에 대한 유전자 구조 예측 프로그램으로는 Glimmer[11]가 가장 널리 사용되고 있다. Coding 및 non-coding region에서의 6-tuple의 출현빈도를 측정해서 coding region을 찾는 GenMark와는 달리 Glimmer에서는 interpolated Markov model을 사용하여 8-tuple 또는 그 이하의 길이를 가지는 oligomer의 출현빈도를 측정해서 유전자 구조 예측에 이용하였다.

진핵생물의 유전자는 원핵생물의 유전자보다 구조가 더 복잡하고 유전체 크기에 비해 유전자의 밀도가 원핵생물보다 훨씬 떨어진다. 원핵생물의 유전자 구조가 Promoter, start codon, coding region, stop codon, non-coding region등으로 이루어진데 비해 진핵생물의 유전자는 cap, polyA와 같이 전사에 관련된 signal이 더 존재하며, coding region도 donor, acceptor signal에 의해 exon과 intron으로 나누어진다.

진핵생물에 대한 유전자 구조 예측은 Caenorhabditis elegans의 유전자 구조 예측에 사용된 gm[13]과 포유동물의 유전자 구조 예측을 연구한 Gelfand[14]에 의해서 시작되었다. 이 두 프로그램은 입력 염기서열로부터 initial exon과 terminal exon을 포함하는 완성된 구조의 유전자를 찾아내고자 하였다.

이 후 hierarchical rule을 이용하여 exon의 가능성이 있는 단편에 대해 순위를 계산하는 모델을 사용한 GeneID[15], neural network과 dynamic programming을 혼합한 GeneParser [16,17], linguistic method를 사용한 GenLang[18], discriminant analysis를 사용한 FGENEH[19], decision tree를 사용한 morgan[20], generalized hidden markov model을 사용한 Genie[21], 그리고 duration hidden markov model을 사용한 GenScan[22]등이 개발되었다.

HMM(Hidden Markov Model)이 응용된 모델에서는 유전자를 구성하는 5'UTRs, 3'UTRs, exons, introns등의 segment를 state로 두고 주어진 DNA 염기서열의 염기들을 그 구조에 포함시킴으로 DNA 염기서열이 형성하는 구조를 밝혀낸다. GenScan에서 이용된 모델인 duration HMM은 HMM이 응용된 형태로서 DNA를 구성하는 각 segment의 길이 분포를 잘 반영할 수 있는 모델이다. 본 연구팀은 EGSP (Eukaryotic Gene Structure Prediction) 프로그램에 GenScan에서 사용된 duration HMM과 signal score 계산방식의 기본 모델을 적용하였고, 각 모듈 및 파라메터 계

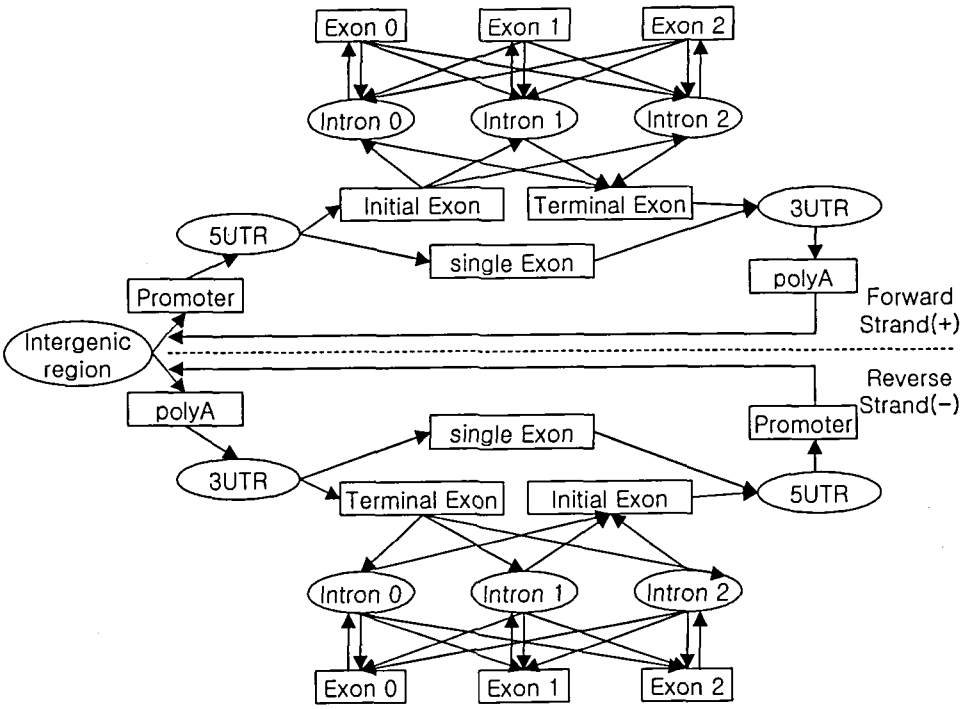


Fig. 1. Duration HMM을 이용한 EGSP의 eukaryotic gene structure prediction model
사각형과 동그라미는 hmm의 state들을 나타내며 각 state들 간의 transition은 오직 화살표 방향으로만 가능하다. Intergenic region을 제외한 모든 state들은 forward와 reverse strand에 따라 구분되어 한 쌍씩 존재한다. Forward와 reverse strand의 transition 방향은 반대방향으로 향한다.

산을 위한 프로그램들을 개발하였다.

EGSP의 학습 데이터와 테스트 데이터로는 human genome의 일부 염색체를 사용하였고, 같은 테스트 데이터에서 수행된 GenScan, GeneID 그리고 Morgan의 결과와 상호 비교하였다.

시스템의 구현과 방법

Duration HMM

HMM은 조건부 확률에 근거해서 바로 이전 상태로부터 현재 상태를 추측해 내는 markov chain의 확장된 모델로서, 음성 인식 분야에 주로 사용되다가 기본적인 통계적인 방법으로 활용되고 있다[23]. HMM은 state들과 symbol들을 기본 요소로 가지고 있으며 관찰되는 symbol들의 출현빈도를 측정해서 조건 확률에 따라 실제에 가장 근사한 state들의 구성을 추측해 내는 모델이다. 1990년 초반부터 motif domain 검색이나 promoter prediction등과 같은 생물학을 위한 연구에서 HMM이 적용되기 시작했다 특히, DNA 염기서열과 같이 연속적이고 반복되는 형태를 가진 구조에서 HMM을 구성하기가 용이하다.

HMM이 효율적인 통계적 모델임에도 불구하고 같은 state가 여러 시간동안 지속될 경우 state들의 지속 횟수에 따른 확률분포를 표현하지 못한다. Duration HMM은 symbol들이 연속해서 같은 state를 지속한 횟수를 표현하기 위한 HMM의 응용된 모델이다. Exon이나 intron과 같은 유전자의 각 구성요소들은 일정한 길이 범위에서 분포하기 때문에 길이의 분포확률은 유전자의 구조를 표현하는데 중요한 정보라고 할 수 있다.

본 논문에서 구현한 duration HMM은 intergenic region과 forward/reverse

strand 별로 각각 promoter, polyA, 5'UTR, 3'UTR, single exon, initial exon, terminal exon, internal exon(phase 0, 1, 2), intron (phase 0, 1, 2), 27개의 hidden state들과, A, T, G, C, 4개의 symbol로 구성 되었다(Fig. 1).

각 state들 간의 상태전이는 Fig. 1의 화살표의 진행방향으로만 일어날 수 있다. 그리고 각 state들 간의 상태전이는 start, stop, donor, acceptor, promoter, polyA, 이렇게 6종류의 signal이 존재하는 위치에서만 가능하다.

학습데이터

이 모델의 구현에 필요한 parameter를 학습시키기 위해서는 많은 종류의 학습데이터가 필요하다. 길이의 분포와 state들의 개시확률, intergenic region의 homogeneous markov matrix를 학습시키기 위해서 Human chromosome 21번을 학습데이터로 사용하였고, homogeneous markov matrix 및 nonhomogeneous markov matrix를 위해 GenBank의 Ref Database Human data와 David Kulp에 의해 1992년에 GenBank release 89로부터 만들어진 dataset을 사용하였다. Promoter는 학습데이터로는 EPD의 Human promoter를 사용하였고, promoter를 제외한 5종류의 signal은 Kulp의 dataset을 이용해서 학습시켰다.

Signal 검색

이 모델에서 state들의 상태전이는 signal의 위치에서만 이루어지므로 signal의 정확한 위치를 찾는 것이 무엇보다 중요하다고 할 수 있다. Signal은 그 메카니즘에 따라 전사관련 signal, 번역관련 signal, splicing signal 이렇게 세 종류로 나눌 수 있다. 그 중에서도 전사관련 signal인 promoter와 splicing signal인 donor, acceptor는 그 위치를 정확히 예측하기 위한 연구가 개별적으로 진행 되고 있을 만큼 생물학적인 중요성이 크다고 할 수 있다.

전사관련 signal인 promoter와 polyA는 유전자의 구조와 intergenic region의 경계가 되는 지점이다. Promoter는 DNA로부터 RNA가 전사되는 정도를 조절하는 중요한 부위로서 다양한 조절인자가 promoter에 영향을 주어서 RNA의 전사를 조절한다. 유전자에 따라서 그 발현을 조절하는 인자가 다르기 때문에 promoter의 구조도 유전자에 따라 다르지만 어느 정도 유사성을 나타내는 부분이 있다. 전사가 시작되는 전사개시부위에 cap site가 존재하고, 전사개시부위의 상류 35-base 부위에 TATA box라는 T, A가 많은 부위가 존재한다. 하지만 TATA box도 현재 밝혀진 전체 promoter의 70% 가량만이 가지고 있기 때문에 이 부위를 이용해 promoter를 검색하는데 많은 어려움이 있다. Cap site와 TATA box 부위를 포함해서 promoter를 검색하기 위해 40 base 길이의 2nd-WAM(Weight Array Matrix)을 이용하였다. 2nd-WAM은 위치 (i-2)의 염기구성 bi-2, 위치 (i-1)의 염기 구성 bi-1에 의해서 i번째의 염기의 조건부 확률 Pi = (bi | bi-1,bi-2)를 구한다. EPD에서 검색한 Human promoter 서열 -39 ~ 0 부위를 학습데이터로 이용해서 promoter 검색에 필요한 matrix를 구성하였다. PolyA(polyadenylation)은 진핵생물의 pre-mRNA 3' 말단에 나타나는 구조로서 일반적으로 AATAAA hexamer의 consensus를 보인다. PolyA를 검색하기위해 GenBank에서 "polyA_signal"로 annotation된 염기서열들을 학습데이터로 이용해서 6 base들에 대한 WMM(Weight matrix model)을 구성하였다.

번역관련 signal인 start signal과 stop signal은 유전자구조에서 coding 부위와 non-coding 부위를 경계 짓는 signal이다. Start codon은 항상 ATG 3 base로 구성되어 있고, stop codon은 TAG, TGA,

Table 1. Acceptor에 대한 각 model의 결과.

TP : Annotation 된 exon중 Prediction 된 exon, FP : Prediction 된 exon중 Annotation 되지 않은 exon
FN : Annotation 된 exon중 Prediction 되지 못한 exon
Sn : Annotation 된 전체 exon에 대한 TP의 비율, Sn : Prediction 된 전체 exon에 대한 TP의 비율

	TP	FP	FN	Sn(TP/Annotated)	Sp(TP/Predicted)
WMM	771	77725	378	0.67	0.009822
WAM	765	98822	382	0.66	0.007682
WWAM	775	93752	372	0.67	0.008199

TAA 이렇게 세 종류의 codon을 보인다. 두 signal 모두 동일한 염기서열을 보이는 다른 부위와의 구분에 큰 특징을 보이지 않기 때문에 정확한 signal의 검색이 어렵다. Start signal의 검색을 위해서는 start codon 이전의 6 base와 이후의 3 base, stop signal의 검색을 위해서는 stop codon 이전의 3base와 이후의 6 base에 대한 WMM을 각각 구성하였다.

Splicing signal인 donor와 acceptor는 exon과 intron의 경계가 되는 signal이다. Donor와 acceptor는 각각 intron의 5' 말단과 3' 말단에 존재하는 부위로 두 signal의 정확한 위치를 찾기 위해 많은 연구가들이 splicing에 대한 연구를 진행하였다. GT 주변에 존재하는 염기들의 구성이 상호의존적이라는 분석에 근거하여 GenScan에서는 MDD(Maximal Dependence Decomposition)와 같이 염기들의 상호의존성을 반영한 모델을 사용하였다. 이 논문에서는 donor의 검색을 위해서 donor의 보존서열인 GT 이전의 3 base와 이후의 4 base에 대해 WMM을 구성하였다. Acceptor signal은 염기들의 상호의존성이 donor와 같이 강하지 않기 때문에 MDD와 같은 모델이 적당하지 않다. Acceptor는 보존서열 AG와 상류에서 약한 상동성을 보이기 때문에 AG 상류 20 base까지 WMM과 WAM 그리고 WWAM(Windowed WAM)을 구성하였다. WWAM의 i번째 값은 i-2, i-1, i, i+1, i+2의 WAM 평균에 의해 구성된다. Acceptor에 대한 세 모델을 비교 분석한 결과 그 차이가 미세한 것으로 나타났다(Table 1). EGSP에서는 AG 상류 20 base에 대해 WMM을 구성하고 acceptor signal을 검색하였다.

State의 길이분포

State의 개수가 L개이고 시간의 길이가 N인 duration HMM에서 Viterbi 알고리즘을 이용해 최적화 된 경로를 찾을 경우 $O(L^2N^3)$의 time complexity를 가진다. 길이가 긴 유전체의 검색에 이러한 time complexity는 적당하지 못하다. 하지만 각 state들의 최소길이와 최대길이에 대한 제한하게 되면 길이가 길어짐에 따라 계산시간이 기하급수적으로 늘어나는 것을 방지할 수 있다.

유전자 구조 예측에서 유전체를 구성하는 state들의 길이 분포는 중요한 정보 중 하나이다. 예를 들어 학습데이터에 의하면 single exon의 최소 길이는 9 base이고 최대 길이는 7400 base이고 9 base 에서 400 base 사이에서는 Gauss distribution에 가까운 분포를 보이는 반면 intron의 경우 exponential distribution을 보인다. EGSP에서는 이러한 분포를 함수로 표현하지 않고 실질적인 분포에 따라 길이분포확률을 적용하였다. 학습데이터를 일정한 길이 단위에 따라 구간별로 나눈 후, 그 구간에 속하는 학습데이터의 개수를 전체 학습데이터로 나누어서 그 구간에 대한 길이분포확률을 할당하였다.

최적의 Gene Structure 구성

EGSP는 입력된 matrix들과 염기 서열로

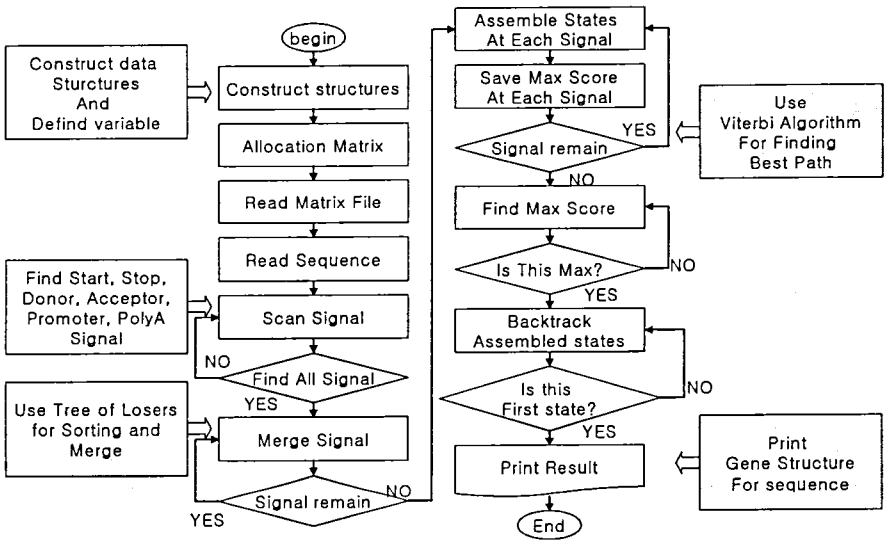


Fig. 2. EGSP의 Eukaryotic gene structure prediction Flowchart

부터 6가지의 signal을 각각 검색하게 되고 일정한 score 이상이 되는 염기서열을 각 signal의 후보로서 저장하게 된다. 각 signal의 후보들이 선택되면 각 signal의 위치에서 가능한 state의 조합을 구성하고 Viterbi 알고리즘을 이용해서 가장 높은 score를 가지는 state의 구성을 찾는다 (Fig. 2).

위치 t에서의 state i일 경우의 score는 아래의 식 (1)과 같이 계산한다.

이 식에서 t는 HMM의 시간 t를 나타내는데, 여기서는 염기서열 내의 base 위치를, T는 염기서열 전체 길이를 의미하고 k는 state i 이전에 올 수 있는 state k를 의미한다. 함수 $F_i(S_{t_{prev},t})$는 t_{prev} 에서 t 까지의 염기서열이 state i에 속할 때 나타나는 확률 값을 반환하고 $L_i(t-t_{prev})$는 state i가 $t-t_{prev}$의 길이를 가질 확률 값을 반환한다. 이 식은 t의 값이 입력 염기서열 전체의 길이인 T에 도달할 때까지 계속되고, T에서 가장 높은 score를 가지는 state i로부터 backtracking을 통해 최적의 gene structure를 구성한다.

결과 및 고찰

1996년 Kulp가 Genie[21]의 구현을 위해 GenBank release 95로부터 만든 dataset에서 학습데이터와 중복되는 부분을 제외시킨 dataset을 테스트 데이터로 이용하였다. 테스트 데이터는 single exon으로 이루어진 유전자 dataset이 208 set, multiple exon으로 이루어진 유전자 dataset이 210 set으로 구성되어있다. EGSP와의 상호 비교를 위한 프로그램으로 GenScan, GeneID, 그리고 Morgan을 사용하였다.

Signal level에서의 결과

Annotation과 prediction의 구분이 명확한 signal인 start, stop, donor, acceptor만을 비교하였다. 이 비교자료는 exon의 종류에 따른 유전자 구조 예측 프로그램들의 정확성

(1)

for $(1 <= t <= T)$

$$r(t,i) = MAX\{\pi_i^* \cdot F_i(S_{1,t})^* L_i(t),\ MAX(t_{prev},k)\{r(t_{prev},k)^* F_k(S_{t_{prev},t})^* L_k(t-t_{prev})\}\}$$

Table 2. Gene structure prediction 프로그램들의 signal에 대한 정확성 비교.

		TP	FP	FN	Sn(TP/Annotated)	Sp(TP/Predicted)
EGSP	start	159	271	256	0.38	0.36
	stop	92	24	326	0.22	0.79
	donor	697	969	441	0.61	0.41
	accept	680	698	469	0.59	0.49
GenScan	start	298	126	116	0.71	0.70
	stop	326	118	92	0.77	0.73
	donor	1014	451	138	0.88	0.69
	accept	1008	455	148	0.87	0.68
GeneID	start	154	102	259	0.37	0.60
	stop	233	79	185	0.55	0.74
	donor	881	329	244	0.78	0.72
	accept	895	351	269	0.76	0.71
Morgan	start	172	235	245	0.41	0.42
	stop	143	264	275	0.34	0.35
	donor	761	1188	413	0.64	0.39
	accept	734	1211	377	0.66	0.37

Table 3. Gene structure prediction 프로그램들의 exon에 대한 정확성 비교.

	Annotated exons			predicted exons		
	# of Exon	%Exac	%Part	# of Exon	%Exact	%Part
EGSP	1560	42	19	1795	36	16
GenScan	1560	78	12	1893	64	10
GeneID	1560	60	17	1512	62	17
Morgan	1560	40	31	2354	26	22

을 나타낸다. EGSP를 비롯한 네 개의 프로그램 모두 start와 stop signal의 정확성이 donor와 acceptor signal에 비해서 떨어지는 것으로 나타났다(Table 2). 이는 유전자 구조 예측 프로그램들이 initial exon과 terminal exon을 찾는데 비교적 어려움을 가지고 있다는 것을 의미한다.

이 결과는 signal을 찾기 위한 모델로 GenScan에서 사용된 MDD(Maximal Dependence Decomposition)가 가장 효과적임을 나타내며, 이러한 모델의 추가를 통해 EGSP가 개선 될 것으로 예상된다.

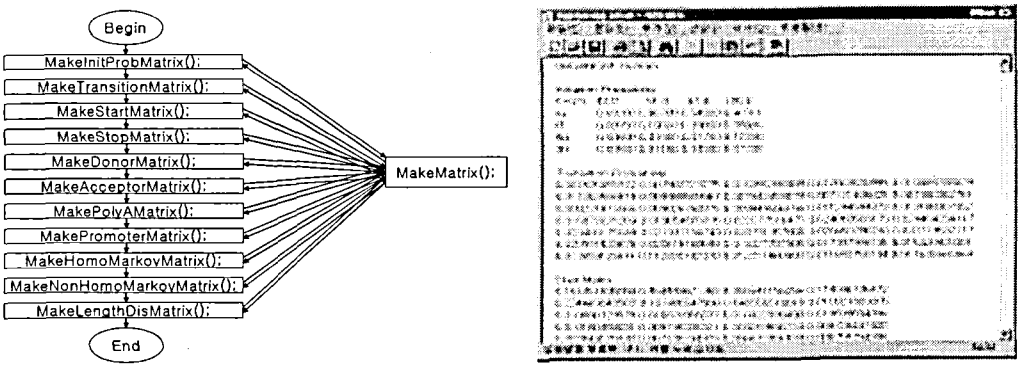
Exon level에서의 결과

Table 3은 네 프로그램들이 prediction 한 exon들을 annotation된 exon들과 비교한 결과를 보여준다. 이 table에서 %Exac는 완전히 일치하는 exon의 비율을 의미하고, %Part는 exon의 양쪽 말단 중 한쪽 말단만이 일치하는 exon의 비율을 의미한다. 비교결과 GenScan이 annotated exon중 78%를 찾아내었고, EGSP는 42%를 찾아내었다. EGSP는 GenScan보다는 낮은 정확성을 보이지만, Morgan에 비해서는 높은 정확성을 보였다.

추후 개선방향

EGPS의 파라메터들은 학습에 필요한 데이터를 독립적인 모듈을 통해서 분석하고 학습한 후 하나의 matrix파일에 기록된다(Fig. 3). 그리고 EGSP는 웹에서 작동하며 사용자로부터 유전자 구조 예측을 수행할 염기서열과 해당 염기서열이 속하는 유전체의 matrix를 입력받은 후 결과를 웹 브라우저를 통해서 보여준다(Fig. 3).

각 signal의 후보 위치에서 최적화된 state

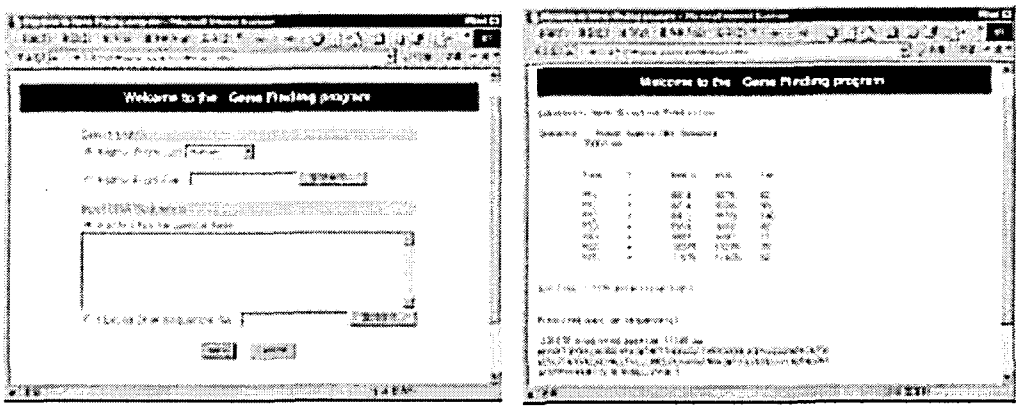


(A) (B)

Fig. 3. EGPS의 matrix 파일을 만들기 위한 학습과정 및 결과

(A) Matrix를 구성하는 각 파라메터들은 별도의 학습과정을 통해 공통된 파일로 기록된다.

(B) 파라메터를 기록한 matrix 파일



(A) (B)

Fig. 4. EGPS의 실행화면 및 결과화면.

(A) EGPS는 웹 브라우저를 통해서 사용자로부터 사용될 matrix의 종류와 유전자 구조예측에 사용될 염기서열을 입력받는다.

(B) 유전자 구조예측이 완료되면 위치정보와 함께 단백질 서열로 변형 했을 때의 서열을 표시한다.

의 구성을 계산하기 위해서 initiation probability, transition probability, length distribution, segment probability, signal probability 이렇게 5가지 계산 모듈에 대해 각각 log odds score 계산 방식을 적용했다. 이 계산 방식은 각 signal의 위치에서 직관적인 비교 값을 제공하지만, 전체적인 score의 비율의 계산에 어려움을 준다. 그에 비해 GenScan에서 사용되는 전체적인 score의 비교방식에서는 서로 다른 계산 모듈에 대한 score의 직관적인 비교가 어렵다. 추후에는 두 방식을 적절하게 조화시켜 나갈 예정이다. 그리고 state들의 segment probability를 계산 할 때 segment내에서의 5th markov score 분포에 따라 state 별로 다른 weight를 둘 필요가 있음을 알게 되었다. 이와 같은 모듈의 추가를 통해 유전자 위치 분석에 조금 더 정확한 예측 결과를 기대할 수 있을 것이다.

참고문헌

[1] J. C. W. Shepherd, Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and ist possible evolutionary justification, *Proc. Natl. Acad. Sci. USA,* 78, 1981,

1596-1600.

[2] J. W. Fickett, Recognition of protein coding regions in DNA sequences, *Nucl. Acids Res.* 10, 1982, 5503-5518.

[3] R. Staden and A. D. McLachlan, Codon preference and its use in identifying protein coding regions in long DNA sequences, *Nucl. Acids Res.* 10, 1982, 141-156.

[4] J. M. Claverie and L. Bougueleret, Heuristic informational analysis of sequences, *Nucl. Acids Res.* 14, 1986, 179-196.

[5] C. J. Michel, New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation, *J. Theor. Biol.* 120, 1986, 223-236.

[6] B. D. Silverman and R. Linsker, A measure of DNA periodicity, *J. Theor. Biol.* 118, 1986, 295-300.

[7] D. G. Arques and C. J. Michel, Periodicities in coding and noncoding regions of the genes, *J. Theor. Biol.* 143, 1990, 307-318.

[8] A. K. Konopka and J. Owens, Complexity charts can be used to map functional domains in DNA, *Genet. Anal. Tech. Appl.* 7, 1990, 35-38.

[9] E. Uberbacher and J. Mural, Locating protein coding regions in human DNA sequences by a multiple sensor-neural network approach, *Proc. Natl. Acad. Sci. USA* 88, 1991, 11261-11265.

[10] M. Borodovsky and J. McIninch, "GENMARK: parallel gene recognition for both DNA strands", *Computer & Chemistry,* 17(2), 1993, 123-134.

[11] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, Microbial gene identification using interpolated Markov models, *Nucl. Acids Res.* 26(2), 1998, 544-548.

[12] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, Improved microbial gene identification with GLIMMER, *Nucl. Acids Res.* 27(23), 1999, 4636-4641.

[13] C. A. Fields, and Soderlund, gm: A practical tool for automating DNA sequence analysis, *Comp. Appl. Biosci.* 6, 1990, 263-270.

[14] M. S. Gelfand and M. A. Roytberg, Prediction of the intron-exon structure by a dynamic programming approach, *BioSystems* 30, 1993, 173-182.

[15] R. Guigo, S. Knudsen, N. Drake, and T. Smith, (1992) Prediction of gene structure, *J. Mol. Biol.* 226, 1992, 141-157.

[16] E. E. Snyder and G. D. Stormo, Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks, *Nucl. Acids Res.* 21, 1993, 607-613.

[17] E. E. Snyder and G. D. Stormo, Identification of protein coding regions in genomic DNA, *J. Mol. Biol.* 248, 1995, 1-18.

[18] S. Dong and D. B. Searls, Gene structure prediction by linguistic methods, *Genomics* 23, 1994, 540-551.

[19] V. V. Solovyev, A. A. Salamov, and C. B. Lawrence, Predicting internal

exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, *Nucl. Acid. Res.* 22, 1994, 5156-5163.

[20] S. Salzberg, A. L. Delcher, K. H. Fasman, and J. Henderson, A Decision Tree System for Finding Genes in DNA, *J. Comp. Biol.* 5(4), 1998, 667-680.

[21] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA, *ISMB-96*.1996.

[22] C. Burge and S. Karlin, Prediction of Complete Gene Structures in Human Genomic DNA, *J. Mol. Biol.* 268, 1997, 78-94.

[23] L. R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proc. IEEE*, 77(2), 1989, 257-285.