ESTs 서열의 생물학적 기능 분류

김태형[1*] 허철구[2] 전여진[3] 김대수[4] 김희수[5]
[1,3,4,] 부산대학교 생물정보학 협동과정
[1,3,4,5] 부산대학교 자연과학대 생물학과
[1,2] 국가 유전체 정보센터
kth2001@pusan.ac.kr

Biological Function Category of ESTs

Tae-Hyung Kim[1*] Cheol-Goo hur[2] Yeo-Jin Jeon[3] Dae-Su Kim[4] Heui-Soo Kim[5]
[1,3,4] Interdisciplinary program of bioinformatics, Pusan National University
[1,3,4,5] Dept of Biology, College of Natural Sciences, Pusan National University
[1,2] Laboratory of Computational Biology, National Bioinformatics Center, Korea Research Institute of Bioscience and Biotechnology, Oun-dong 52, Yu-Sung, Daejon, 305-333 Korea

요 약

실험실에서 대량으로 그리고 빠른 속도로 결정되고 있는 ESTs를 이용하여 유전자를 동정하고 생물학적 기능분류를 돕기 위한 몇 가지 유용한 방법을 제안하게 되었다. 첫번째로 single linkage 알고리즘을 이용하여 ESTs clustering하는 BEC (Blast ESTs clustering) 프로그램을 개발하여 동일한 유전자 유래 ESTs를 밝혀내었으며 virus 또는 small RNA의 유래된 내재성 반복서열이 포함되어 발현된 ESTs을 분류하였다. 다음으로 순차적인 생물학적 기능 분류 하기 위하여 먼저 Riken의 Fantom full-length mouse cDNA와 NCBI의 Refseq mouse mRNA을 이용하여 이미 밝혀진 유전자를 확인하는 작업을 수행하였다. GO (Gene Ontology)의 3가지 큰 분류에 따라 hierarchical하게 분류를 하였다. No-hit된 ESTs는 단백질로 번역된 후 유사성 정도에 따른 3단계로 나누어 분류하였다. 마지막까지도 Unknown으로 분류된 ESTs에 대해서는 도메인 및 모티프 검색을 통하여 잠재적 코딩영역에 대한 분석과 함께 기능별 분류를 하였다. 이렇게 제안된 여러 분석작업을 통해 우리는 ESTs에 대한 clustering, quality 검사 및 여러 단계에 따른 생물학적 기능분류를 종합적으로 할 수가 있었다.

1. 서 론

현재 ESTs는 지놈 서열 다음으로 많은 양의 데이터가 공개 되어 있으며 시스테믹한 유전자 연구를 함에 있어 접근을 용이하게 하는 장점이 있다. 여러 다양한 조직에서 발현되는 전사체로부터 얻어진 cDNA를 이용하여 대량으로 랜덤 하게 얻어진 이들 ESTs를 이용함으로 해서 잠재적 코딩영역 및 유전자의 발현 유무 그리고 유전자 구조를 밝혀 내는데 있어서 비교적 적은 비용과 시간을 들여 연구할 수 있는 장점을 가지고 있다[1-4].

하지만 ESTs 자체는 단 한번의 서열 결정을 통해 얻어짐으로 해서 비교적 높은 에러비율(1/100bp)을 가지고 있으며[5,6], 세포내 유전자에서 발현된 전사체가 아닌 인트론서열[7] 및 내재서열[8](virus like elements, rRNA like elements)과 같은 지노믹 서열이 포함되어 cDNA가 만들어지고 이들이 ESTs 데이터로써 서열결정이 되는 경우가 흔히 나타나고 있으며 ESTs 그 자체가 조각난 어떠한 특성으로 인해 이들을 원래 상태의 유전자구조로 복원 하는데 있어서 문제점이 완벽히 해결되지 않은 상태이며 수 많은 전처리 작업과 작업계획과 비교적 거대한 계산 시스템을 요구 하고 있다는 것이 문제점으로 지적되고 있다.

그래서 본 작업은 ESTs의 고유한 문제점들을 극복하고

장점을 살리기 위해 ESTs를 클러스터링을 통해 같은 유래 ESTs를 분류하고, 내재성 반복서열에 대한 검출과 분류작업을 하였으며 순차적으로 ESTs의 생물학적 기능분류 하는 방법을 제안하게 되었다. 이들 기능분류에 GO node[9]들을 사용하여 계층적으로 즉 대분류, 중분류, 소분류에 따른 ESTs의 생물학적 기능분류에 이용하였다. 여기 본론에서 보여주고자 하는 것은 주어진 마우스의 EST 3,000개를 예로 제시한 방법들을 통해 생물학적 기능 분류에 대해 설명하고자 한다.

방법 및 결과

1. ESTs 클러스터링 작업

사용한 ESTs 클러스터링 방법으로 minimal linkage (transitive closure, single linkage)[10]를 채택하였으며 BLAST를 이용하여 어느 역치 (threshold)이상의 점수를 가지는 ESTs들은 자신의 cluster에 합쳐나가며 또 다음 cluster와 비교하여 역치 이상이면 이들 cluster들을 합쳐나갔다. 다음은 구현된 알고리즘에 대한 설명이다.

초기 상태: 각 서열은 자신의 cluster를 가지고 있다. (Si 는 cluster i 이다 or Ci = i).

첫 번째 반복: 제일 먼저 나오는 entry S0는 query로 선택되었다. 나머지 서열들은 Si ($1 \leq i < N$)

MERGE (cluster C0) ← *(cluster Ci) if d2 (S0, Si) < THRESHHOLD.*

두 번째 반복: 두번째 서열이 S1 query가 되며 동시에 cluster가 되었다. 나머지 서열들은 Si ($2 \leq i < N$)

MERGE (cluster *C1* ← cluster *Ci*) if *d2* (*S1, Si*) < THRESHHOLD.

(*k*) 번째 반복: 전체적으로 ($k - 1$)번 반복하게 된다. k-1을 *Sk* query로 선택하게 되며 또한 자신의 cluster를 가지게 된다. 나머지 서열들은 ($k + 1 \leq i < N$)

MERGE (cluster *Ck* ← cluster *Ci*) if d2 (*Sk, Si*) < THRESHHOLD.

이러한 알고리즘을 구현하여 BEC(Blast ESTs clustering)이라는 ESTs 클러스터링 하는 GUI프로그램을 파이썬으로 개발하였다. 이 프로그램은 수십만 개의 ESTs가 주어지더라도 자동적으로 formatdb를 형성하며 BLAST의 빠른 서열정열 방법을 이용하여 클러스터링을 하는데 이용하였으며 순서는 UniGene의 ESTs 클러스터링과 유사한 방식을 채택하였다. 이전연구로 프로그램에 대한 성능평가를 위해 우리는 dbEST Library에서 마우스의 bone 조직의 12,709 개 ESTs에 대한 BEC프로그램으로 클러스터링을 해 본봐 bit score 200점에서의 클러스터된 ESTs의 결과와 UniGene 마우스 bone조직의 ESTs의 클러스터링 결과와 아주 흡사함을 알 수가 있었다. CPU 700MHZ, 메모리 370MB 환경의 PC급 컴퓨터 1대에서 작업은 이루어 졌으며 이들 대량의 ESTs에 대한 클러스터링 결과를 비교적 빠르게 얻어 낼 수가 있었다.

Figure 1. python으로 개발된 BEC GUI프로그램.

다음 표는 minimal linkage 알고리즘으로 개발된 BEC 프로그램으로 mouse 3000개 ESTs에 대한 여러 threshold (100~500)에 따른 ESTs가 서로 관계를 가지고 클러스터되는 분포를 보여주고 잇다. 그리고 각각 클러스터에 해당하는 ESTs의 합은 어느 정도인지 그리고 전혀 ESTs간에 관련성이 없어 singleton ESTs로 분류되는 것들은 얼마나 되는지에 대한 분포를 보여주고 있다.

Table 1. Score별 ESTs clustering

Threshold	Clusters	Total ESTs of clusters	Singleton ESTs
100	156	406	2,594
150	154	388	2,612
200	149	363	2,637
250	133	323	2,677
300	122	295	2,705
350	99	239	2,761
400	88	211	2,789
450	73	176	2,824

500	65	157	2,843

2. 내재 반복 서열들에 대한 동정 작업

오랜 마우스의 진화 동안 바이러스 또는 tRNA 및 rRNA의 전사체가 지놈 내 유입을 통해 현재까지도 세포내 유전자 (Cellular Gene)와 함께 발현을 일으키고 있으며 ESTs를 얻기 위한 cDNA 작성시 이들 서열이 포함되어 ESTs서열로써 실제 세포 내 유전자와 함께 얻어지고 있다[11]. 이들은 세포 내에서 기능하는 유전자의 발현을 연구하기 위해서는 방해의 요소가 되기도 하기 때문에 이들을 ESTs 내부에 있어서 반복 서열들을 검출할 수 있도록 하였다. 또한 이러한 반복서열이 포함된 전사체는 직, 간접적으로 인간의 질병 및 생물학적 기능을 담당하고 있음으로 해서 중요한 연구 대상이 되기도 한다[12]. 우리는 이들을 동정하기 위하여 Repeatmasker[13]을 이용하여 마우스 특이적인 반복 서열들을 동정하였다. 이때 비교대상의 서열들은 Repbase[14] 라이버러리를 이용하였다. Table 2에서의 분석결과를 통해 이들 3,000개 마우스 ESTs안에서 628개의 elements가 SINEs (B1s, B2-B4, IDs, MIRs), LINEs (LINE1, LINE2), LTR elements (MaLRs, ERV_classI, ERV_classII)와 같은 Retrotransposon과 DNA elements (MER1_type, MER2_type) 과 같은 DNA transposon, 그리고 다수의 Small RNA, tandem repeat, Low complexity가 검출되었다. 여기서 잠재적으로 발현가능성이 있는 SINEs, LINEs, LTR elements의 경우 repeat elements가 포함된 284개의 ESTs가 검출되었다.

Table 2. 3,000개 ESTs안에 분포되어 있는 repeat서열

```
file name: mouse_3000
sequences:                3000
total length:       1337959 bp  (1337446 bp excl N-runs)
GC level:             47.33 %
bases masked:         70809 bp (  5.29 %)
=================================================
                number of      length   percentage
                elements*    occupied  of sequence
-------------------------------------------------
SINEs:                233       28074 bp      2.10 %
      B1s             116       13081 bp      0.98 %
      B2-B4            88       12481 bp      0.93 %
      IDs              17        1284 bp      0.10 %
      MIRs             12        1228 bp      0.09 %

LINEs:                 32        6573 bp      0.49 %
      LINE1            27        6126 bp      0.46 %
      LINE2             5         447 bp      0.03 %
      L3/CR1            0           0 bp      0.00 %

LTR elements:          64       15427 bp      1.15 %
      MaLRs            24        4041 bp      0.30 %
      ERVL              0           0 bp      0.00 %
      ERV_classI        9        2156 bp      0.16 %
      ERV_classII      25        8386 bp      0.63 %

DNA elements:          11        1471 bp      0.11 %
      MER1_type         8         944 bp      0.07 %
      MER2_type         2         362 bp      0.03 %

Unclassified:          2         634 bp      0.05 %

Total interspersed repeats:     52179 bp      3.90 %

Small RNA:             4         273 bp      0.02 %

Satellites:            1          73 bp      0.01 %
Simple repeats:      213        8908 bp      0.67 %
Low complexity:      220        9411 bp      0.70 %
=================================================
```

여기서 실제로 Retrotransposons이 ESTs로 발현될 가능성이 높은 것으로 보이는 300bp이상 ESTs서열과 반복서열이 매칭된 결과를 Table 3. 에서 21개의 ESTs 와 반복서열에 대한 관계를 보여주고 있다. 결과에서 보듯이 L1, ERVK의 Retrotransposons이 마우스 ESTs에서 포함되어 발현됨을 알 수가 있었다. 이 결과로 보아 이들 외래 유전자는 강력하게 그리고 흔하게 발현되어지고 있음을 유추해볼 수가 있었다.

Table. 3. ESTs 와 300bp 이상 매칭된 반복 서열

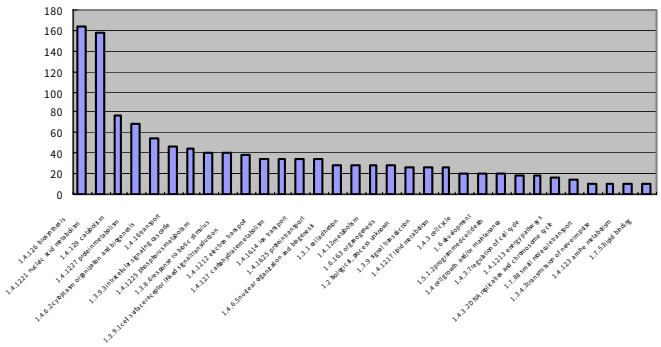
EST	Repeat	Family	Length	Direction
313_seq	MTA	LTR/MaLR	325	C
333_seq	IAPEz-int	LTR/ERVK	448	+
387_seq	Lx2	LINE/L1	398	+
404_seq	IAPEz-int	LTR/ERVK	480	+
1015_seq	L1	LINE/L1	638	+
1692_seq	Lx2	LINE/L1	355	C
1933_seq	ETnERV3	LTR/ERVK	830	+
1934_seq	L1	LINE/L1	396	C
1946_seq	Lx4	LINE/L1	438	C
2019_seq	ORR1A1	LTR/MaLR	346	+
2095_seq	IAPEz-int	LTR/ERVK	376	+
2100_seq	L1	LINE/L1	361	C
2134_seq	IAPEz-int	LTR/ERVK	878	+
2162_seq	L1	LINE/L1	420	+
2178_seq	IAP-d	LTR/ERVK	448	+
2414_seq	RMER17A2	LTR/ERVK	577	+
2429_seq	MTD	LTR/MaLR	409	C
2440_seq	RMER4B	LTR/ERVK	405	+
2509_seq	MMETn-int	LTR/ERVK	382	+
2670_seq	MER34B	LTR/ERV1	429	+
2964_seq	MMETn-int	LTR/ERVK	475	+

3. 순차적인 ESTs 기능분류작업

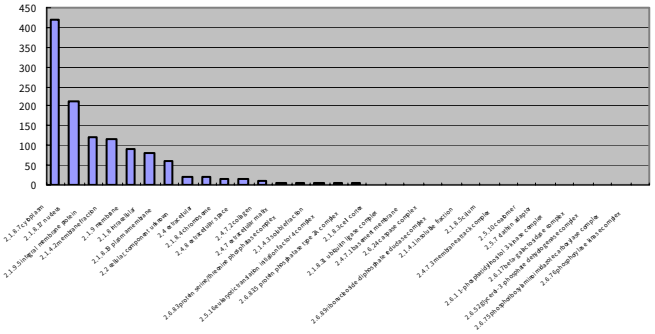
이들 ESTs를 RIKEN의 Fantom 마우스 cDNA와 NCBI의 Refseq 마우스 mRNA (NM_XXXXXX)를 이용하여 ESTs을 기존의 잘 알려진 유전자에 분류를 시켰다.

MGI annotation data 및 계층적으로 다시 가공한 GO정보를 통해 유전자 생산물에 대한 기능 분류체계인 Gene Ontology의 3가지 GO 카테고리 (biological process, cellular component, molecular function)로 나누었다. 그리고 Figure 2. 결과에서는 각 대분류 카테고리에서 3번째 중분류에 해당하는 카테고리까지 분포를 확인하였다. 이때 3가지 각각 대분류에 따른 가장 높은 기능분류 분포를 보여주는 중분류로는 다음과 같다. Biological process에서는 1.4.12.6 (Biosynthesis), 1.4.12.21 (nucleic acid metabolism), 1.4.12.9 (catabolism)로 Cellular component에서는 2.1.8.7 (cytoplasm), 2.1.8.17 (nucleus), 2.1.9.5 (integral membrane protein), 마지막으로 Molecular function에서는 3.12.25.41 (protein kinase), 3.6.18.3 (purine nucleotide binding), 3.6.17.1 (DNA binding)로 ESTs가 기능하는 유전자로써 가장 많이 확인되어 분류가 되었다. 여기서 사용된 GO 데이터는 MGI[15]에서 마우스의 단백질들을 GO로 분류 및 해석한 데이터를 사용하였으며 GO node에 따라 계층적인 구조를 기능별 단계별로 그리고 연구자들이 보기 쉽게 biological process는 1번, cellular component는 2번 molecular function은 3번으로 이러한 식으로 3가지 대분류에서 시작해서 1.xx.xx.… 와 같이 중분류, 소분류로 계속 기능이 분류되어 만들어 보여주고 있다. 이렇게 GO 정보를 이용 함으로 해서 수많은 ESTs의 하나하나에 대한 기능을 계층적으로 확인할 수가 있었으며 마우스 3,000개 ESTs에 대한 전체적인 생물학적 기능 분포를 확인할 수가 있었다.

A. Biological process



B. Cellular component



C. Molecular function

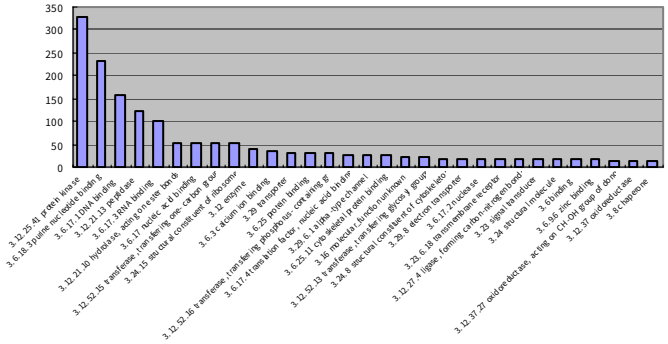


Figure 2. 3가지 GO(Gene Ontology) category 분류.

이미 기능이 알려진 유전자로 분류된 ESTs에 대해서는 앞의 작업과 같이 GO 카테고리에 분류를 하였다. 다음 순서로 이들 알려진 유전자와의 No-hit된 ESTs에 대해서는 단백질 서열과의 상동성 검색을 하였다. 이를 위해 BLASTX를 사용하여 ESTs를 단백질로 번역하여 NR (Non-redundant) 단백질서열과의 비교를 하였다. 이때 100bp이상에서 90%이상의 identity를 가지는 ESTs는 기능적으로 이들 단백질들과 상동성을 가지는 것으로 보았으며 90%~70% identity를 가지는 ESTs는 이들 단백질과 기능적으로 유사할 것으로 그리고 마지막으로 70%~40% identity를 가지는 ESTs는 낮은 identity이지만 기능적으로 진화적 관련이 있을 것으로 보았다. 이러한 3단계에 따른 단백질과의 상동성 검색에서도 생물학적 기능이 전혀 알려지지 않은 ESTs에 대해서는 Interproscan 프로그램[16]을 이용하여 단백질의 기능적으로 중요한 모티프 및 도메인을 검색하였다. 이 프로그램에 내장되어 있으며 활용된 툴들은 단백질 Profile scan과 Fingerprint

scan이였다. 이러한 단백질 모티프 및 도메인 검색을 수행함으로 해서 ESTs에 있어서 이미 알려진 유전자에 대한 분류는 안되었지만 잠재적인 코딩능력의 여부와 새로운 기능을 가진 유전자 검출에 대한 가능성에 대해 제시하고 있다. 마우스 3,000개 ESTs에 대한 기능 분류는 아래의 Figure 3.과 같이 순차적으로 진행되었다.

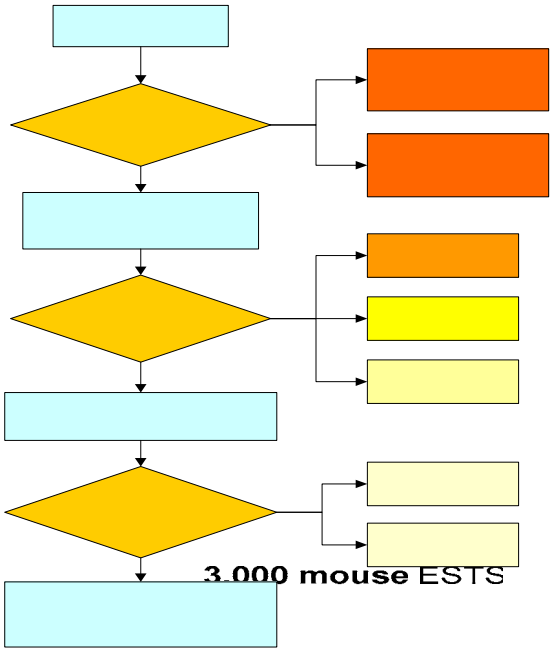


Figure 3. ESTs의 생물학적 기능 분류에 대한 도식도.

결론 및 토의

이들 3,000개 마우스 ESTs에 대한 생물학적 분류 작업을 위하여 제시한 여러 방법들을 이용함으로 해서 ESTs 분석에 있어서 보다 신속하게 그리고 이들과 관련된 유전자를 찾고 생물학적 기능에 대한 분류를 하는데 있어 종합적이고도 효과적인 수행할 수 있을 것으로 보인다. 우리는 minimum linkage 방법을 통해 ESTs를 클러스터링 함으로 해서 유전자의 일부분이 발현되어 결정된 이들 ESTs 서열간의 최소한 중복되는 관계를 통하여 모두 한cluster로 묶음으로 해서 유전자 동정, 구조 및 다른 여러 형태로 발현된 유전자를 찾아 낼 수 있는 정보로 이들을 이용할 수가 있을 것이다. ESTs의 내재된 종 특이적인 반복 서열에 대한 분석을 통해 우리는 이들 ESTs에 대한 반복서열에 대한 내재성에 대한 유무를 확인함으로 해서 이들 ESTs를 외래유래 유전자를 제외한 세포내 발현된 유전자 연구를 위한 분석에 이용할 때 반복서열이 내제된 서열에 대한 제외를 통해 좀더 확실한 유전자 구조 분석 및 기능분류에 다가갈 수 있을 것으로 보아 진다. 그리고 더 나아가 외래 유래 유전자의 강한 발현의 동정을 통해 우리는 전사체에 있어서의 Alternative splicing및 이들 반복서열에 의한 엑손화 현상에 대한 분석을 용이하게 할 수가 있을 것으로 보아 진다. 이러한 반복서열에 대한 분석은 내재되어 있는 반복서열에 대한 ESTs의 문제점을 최소화하는데 반면 연구할 가치가 있는 반복서열에 대한 분석을 가능케 하고 있다. GO에 대한 생물학적 대분류 및 중분류에 해당하는 ESTs에 대한 기능의 분포를 3가지 카테고리에 따라 확인할 수가 있음으로 해서 ESTs 각각의 기능에 대한 폭 넓은 이해를 도울 수 있을 것이다. 그리고 각기 다른 조직에 따른 다른 상황에서의 발현되는 전사체에 대한 기능 분류를 함에 있어 GO를 대분류, 중분류, 소분류로 나누어 계층적인 분류를 하는 것은 유전자 발현 양상을 연구하는 여러 분야의 수많은 연구자에게 유용할 것이다. 나아가 이미 기능이 알려진 유전자가 아닌 그리고 서열에 있어서 유사성은 낮지만 진화적으로 기능적으로 상동성을 가지고 있는 단백질에 대한 동정을 통해 새로운 분류체계 및 모티프 및 도메인 검색을 통해 잠재적인 또한 새로운 기능을 가진 유전자 발굴에 이들 ESTs에 대한 생물학적 분류는 크게 유용하게 활용되어 질것으로 보인다.

참고문헌

1. Adams,M.D., Dubnick,M., Kerlavage,A.R., Moreno,R., Kelley,J.M., Utterback,T.R., Nagle,J.W., Fields,C. and Venter,J.C. Sequence identification of 2,375 human brain genes. *Nature*, **355**, 632–634. 1992.

2. Adams,M.D., Kerlavage,A.R., Fields,C. and Venter,J.C. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nature Genet.*, **4**, 256–267. 1993.

3. Boguski,M.S. The turning point in genome research. *Trends Biochem. Sci.*, **20**, 295–296. 1995.

4. Marra,M.A., Hillier,L. and Waterston,R.H. Expressed sequence tags—ESTablishing bridges between genomes. *Trends Genet.,* **14**, 4–7. 1998.

5. Ewing,B. and Green,P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.,* **8**, 186–194. 1998.

6. Hillier,L.D., Lennon,G., Becker,M., Bonaldo,M.F., Chiapelli,B., Chissoe,S., Dietrich,N., DuBuque,T., Favello,A., Gish,W. *et al.* Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.,* **6**, 807–828. 1996.

7. Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P. and Mattick,J.S.ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet.,* **24**, 340–341. 2000.

8. Weber,G., Shendure,J., Tanenbaum,D.M., Church,G.M. and Meyerson,M Identification of foreign gene sequences by transcript filtering against the human genome. *Nature Genet.,* **30**, 141–142. 2002.

9. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight, S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M. and Sherlock,G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25** 25-29. 2000.

10. Miller,R.T., Christoffels,A.G., Gopalakrishnan,C., Burke,J., Ptitsyn,A.A., Broveak,T.R. and Hide,W.A. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.,* **9**, 1143-1155. 1999.

11. Rotem Sorek, Gil Ast and Dan Graur. Alu-containing exons are alternatively spliced. *Genome Res.* **12**, 1060-1067. 2002.

12. Kim,H.S., Crow,T.J. and Hyun,B.H. Assignment of the endogenous retrovirus HERV-R (ERV3) to human chromosome 7q11.2 by radiation hybrid mapping. *Cytogenet. Cell Genet.* **89**, 10. 2000.

13. Smit, A. and P. Green. RepeatMasker at *http://ftp.genome.washington.edu/RM/RepeatMasker.html.* 1991.

14. Jurka,J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418-420. 2000.

15. Jackson Laboratory. GO association data at ftp://ftp.informatics.jax.org/pub/reports/gene_association.mgi. 2002.

16. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. The EBI SRS server-recent developments. *Bioinformatics,* **18** 368-373. 2002.