유전체 염기서열의 base-composition에 대한 연구

정철희[1] 윤경오[3] 최진영[1] 박현석[2]
고려대학교 컴퓨터학과[1]
세종대학교 컴퓨터공학과[2]
(주)마크로젠[3]
{cholhee, choi}@formal.korea.ac.kr
hspark@cs.sejong.ac.kr
yoonko@macrogen.com

Study on base-compositions of biological sequences

Chol-Hee Jung[1,3] Kyong-Oh Yoon[2] Jin-Young Choi[1] Hyun-Seok Park[2,3]
Dept. of Computer Science & Engineering, Korea University[1]
Dept. of Computer Science & Engineering, Sejong University[2]
Macrogen[3]

요 약

생물체가 생명을 영위하기 위해 수행하는 모든 기능들에 대한 정보는 각 개체가 가지고 있는 유전체에 들어있다. 그런데 각 생물체마다, 심지어는 한 생물체의 서로 다른 염색체마다 그 전체 염기서열에서의 base-composition 은 같지 않고, 또한 이 구성비에는 일정한 특징이 있다. 따라서 이 논문에서는 각 생물체들의 전체 염기서열을 구성하는 염기의 구성비에 대해 조사하고 비교해 보고자 한다.

1. 서론

생물체의 모든 유전 정보를 담고 있는 유전체는 처음부터 끝까지 A, C, G, T 라는 네 개의 염기로 구성된다. 이들 중 A 는 T 와, C 는 G 와 쌍을 이루며 2 중 나선 구조를 만든다. 왓슨-크릭 쌍이라고 부르는 A-T, C-G 쌍을 이용하면 2 중 나선 구조를 이루는 한 가닥의 DNA 로 나머지 한 가닥의 DNA 도 정확히 알아낼 수 있다. 따라서 유전체 전체의 염기서열을 표시할 때는 서로 쌍을 이루는 두 가닥 중 한 가닥의 염기 서열만으로 나타내는데, 이렇게 한 가닥의 염기 서열로 나타낸 전체 유전체의 염기 서열은 {A, C, G, T}를 알파벳으로 사용하는 하나의 긴 문자열로 볼 수 있다. 유전체의 염기 서열을 나타낸 문자열이 4 개의 문자만을 사용한 것이라고 했을 때, base-composition, 즉 각 문자들이 전체 문자열에서 차지하는 비율이 어느 정도인가를 생각해 볼 수 있다.

지금까지 생물학계에서는 유전체 염기서열에서의 GC-rich region(염기서열 중 G 나 C 가 많이 나타나는 영역)에 대한 연구를 주로 진행하였다. 이는 GC-rich region 이 생물학적으로 중요한 의미를 가지고 있기 때문이다. 예를 들어 미생물의 경우 동일 유전자에서의 GC-content(염기서열 중 G 나 C 가 차지하는 비율)는 미생물들 사이의 진화적 거리를 예측하는 자료로 사용된다[1][5]. 이러한 이유들 때문에 미생물 유전체의 전체 염기서열을 밝혀내는 작업을 하는 경우, 마지막 단계에서 GC-content 를 조사하는 작업을 하게 된다.

그렇지만, 각 염기값들 자체의 비율이 어느 정도인가에 대한 연구는 거의 이루어지지 않았는데, 각각의 염기값들은 statictical bias 가 없는 경우 전체 문자열에서 25%의 확률로 나타날 것이지만, 실제 생물체 유전체의 염기 서열에서는 각 염기값들이 25%의 비율로 나타나지 않는다. 이는 생물체 유전체의 염기서열에는 statictical bias 가 존재한다는 것을 보여준다. 이에 본 저자들은 이 논문에서 33 개의 미생물 유전체 염기 서열과 인간 EST 염기 서열[3]에서의 base-composition 을 조사하고 비교, 분석하였다.

2. base-composition 조사 알고리즘 및 구현

생물체 유전체를 구성하는 염기 서열에서 A, C, G, T 의 비율은 다음과 같은 간단한 알고리즘을 구현한 프로그램을 통해 조사할 수 있다[표 1].

[표 1] 염기 값의 빈도를 구하는 알고리즘

염기 값의 빈도를 구하는 알고리즘.	
입력	seq ← 유전체의 염기서열, k ← 분포를 알고자 하는 염기 서열의 개수
출력	유전체 염기 서열을 구성하는 각 염기 값의 비율
	dic : 하나 혹은 여러 개의 염기 값을 키로 하는 해쉬 구조체. i ← 0 : 입력 염기 서열의 인덱스 1) **bases** ← 입력 염기서열의 i 번째로부터 n 개의 염기 값 2) **dic** 에 **bases** 라는 key 값을 가진 항목이 존재 하면 그 항목의 값을 1 증가. 존재하지 않으면 bases 라는 값을 키값으로 하는 항목을 추가하고 그 항목에 1 을 대입. 3) i ← i+1 4) i + n 이 입력 염기서열의 길이보다 크지 않으면 1)로, 보다 크면 5)로 이동 5) dic 에 저장된 모든 항목들을 그 키값과 함께 출력.

위 알고리즘에서 키값이 될만한 것이 A,C,G,T 네 개 뿐인데도 굳이 해쉬 구조체를 사용한 것은 전체 염기 서열에서 염기값 하나 뿐만 아니라 일정 개수의 연속적인 염기값들의 분포를 알기 위한 작업에도 동일한 방법으로 사용될 수 있도록 더욱 일반적인 알고리즘을 작성하기 위해서다. 실제로 세 개(codon)나 여섯 개(dicodon)의 연속적인 염기 값들의 빈도에 대한 데이터[4]는 염기 서열에 숨어 있는 생물학적인 문법을 밝혀내기 위한 작업의 기초자료로 매우 중요하게 사용될 수 있으며 이는 본 저자들의 향후 연구 과제중의 하나다.

위 알고리즘은 인터프리터 언어인 Python 을 이용하여 구현하였다. 위 알고리즘에서 각 염기값들의 빈도수를 저장하는 해쉬 구조체 'dic'은 Python 이 기본적으로 제공하는 해쉬 구조체인 Dictionary 라는 class 를 이용하였다.

3. 염기값의 빈도 수 조사

이 논문에서 저자들이 중점적으로 알아내고자 하는 데이터는 염기값 하나의 빈도이다. 따라서 위 알고리즘을 구현한 프로그램의 입력 중 k 값은 1 로 하였다.

다음 [표 2]는 32 가지 미생물체 유전체의 전체 염기 서열에서 각각의 A,C,G,T 의 빈도를 조사한 결과이다.

[표 2] 미생물 유전체의 base-compositions

Genome name	G%	C%	A%	T%
Aful	24.3758	24.2058	25.8032	25.6152
Aquae	21.7941	21.6820	28.4129	28.1109
Bbur	14.2362	14.3585	35.4766	35.9287
Bsub	21.7097	21.8080	28.1805	28.3018
Buch	13.2292	13.0833	37.0734	36.6142
Cjej	15.2044	15.3442	34.8279	34.6235
Cpneu	20.2600	20.3177	29.8515	29.5708
CpneuA	20.3153	20.2596	29.5734	29.8517
CpneuJ	20.2577	20.3227	29.8592	29.5604
Ctra	20.6619	20.6454	29.4211	29.2716
Dra1	33.4707	33.5412	16.4818	16.5063
Dra2	33.3586	33.3302	16.9550	16.3562
Ecoli	25.3658	25.4231	24.6191	24.5920
Hinf	18.9853	19.1650	31.0173	30.8325
Hpyl	19.2630	19.6112	30.3028	30.8231
Hpyl99	19.4926	19.6961	30.3264	30.4849
Mgen	15.9138	15.7780	34.5720	33.7362
Mjan	15.8907	15.5358	34.4411	34.1324
MpneuM	19.9561	20.0520	29.4662	30.5258
Mthe	24.8131	24.7308	25.0853	25.3708
Mtub	32.7461	32.8680	17.1951	17.1908
NmenA	25.9163	25.8928	23.9964	24.1944
NmenB	25.9713	25.5566	24.2042	24.2679
Paer	32.9901	33.5656	16.8593	16.5850
Pyro	22.2807	22.4332	27.5824	27.7037
Rpxx	14.6207	14.3796	35.3743	35.6254
Synecho	23.8930	23.8273	26.0913	26.1885
Tmar	23.4829	22.7648	26.9698	26.7825
Tpal	26.5684	26.2067	23.5420	23.6828
Uure	12.9427	12.5570	37.2969	37.2034
Vcho1	23.9413	23.7540	25.9778	26.3268
Vcho2	23.6483	23.2655	26.4852	26.6011
Xfas	27.7305	24.9436	22.5450	24.7809
평균	21.6261	21.4972	26.9372	26.9983

위 표에서와 같이 A 와 T, C 와 G 의 비율이 거의 같게 나온다는 것은 전체 유전체에서의 GC-content 부분에서는 G 나 C 가 거의 1/2 의 확률로 나타나고, A 와 T 의 경우에서도 마찬가지라는 것을 의미한다.

유전체 전체 염기 서열에서 A 와 T, C 와 G 의 비율이 비슷한 것은 미생물의 경우만이 아니다. 전체 염기 서열이 거의 정확하게 밝혀진 인간 염색체 21 번 22 번

의 경우도 마찬가지다[표 3].

[표 3] 인간 염색체의 base-compositions

염색체	G%	C%	A%	T%
21 번	20.4240	20.4619	29.6601	29.4540
22 번	23.8996	23.9176	26.1598	26.0229

그러나 유전체 전체 염기서열 중 유전자 부분 만의 염기서열의 경우는 그 비율이 유전자들마다 다르다[표 4].

[표 4] 인간 유전자의 base-compositions(일부)

UniGene #	G%	C%	A%	T%
...	...	...	...	...
ug=Hs.315	18.2061	44.0267	26.1260	11.6412
ug=Hs.240	18.8415	15.3565	38.8924	26.9096
ug=Hs.62	19.3671	18.5443	34.0823	28.0063
ug=Hs.51	19.7827	18.0830	29.2561	32.8782
ug=Hs.2	19.9060	19.6708	31.5047	28.9185
ug=Hs.30	19.9249	17.7527	33.0652	29.2572
ug=Hs.311	20.1220	16.8237	31.4024	31.6519
ug=Hs.149	20.2503	26.6970	28.5552	24.4975
ug=Hs.313	20.2756	20.2100	31.8241	27.6903
ug=Hs.214	20.7845	21.1931	29.0929	28.9294
ug=Hs.316	21.1705	19.2678	30.1060	29.4557
ug=Hs.120	21.7181	22.0811	26.7998	29.4011
ug=Hs.242	21.8094	25.4604	23.7157	29.0145
ug=Hs.315	18.2061	44.0267	26.1260	11.6412
...	...	...	...	...

[표 4]는 NCBI(National Center for Biotechnology Information)에서 제공하는 염기서열 데이터베이스 중 UniGene[2] 항목에 있는 ' Hs.seq.uniq'라는 데이터베이스 파일에 있는 염기서열들에서 A,C,G,T 의 비율을 조사한 것 중 일부이다. ' ug' 는 UniGene 에서의 염기서열 번호이다. 앞서 말했듯이 ug 가 다르면 A,C,G,T 의 비율도 다르다.

그렇지만 Hs.seq.uniq 파일에 저장되어 있는 총 85,903 개 염기서열들 전체에서 A,C,G,T 의 비율을 조사해 보면 A-T, C-G 의 비율이 여전히 비슷하게 나오는 것을 알 수 있다[표 5].

[표 5]

	G%	C%	A%	T%
Total	23.2724	23.003	27.2188	26.506

다음 [표 6]은 본 저자들이 염기서열 분석을 끝마친 *Zymomonas mobilis* 라는 미생물체의 전체 염기서열에서 A,C,G,T 의 비율을 조사한 결과이다.

[표 6] *Z.mobilis* (draft)의 base-composition

	G%	C%	A%	T%
Z.mobilis, draft	23.1892	23.0240	26.8181	26.9687

위 표를 볼 때, *Z.mobilis* 에서도 base-composition 이 다른 미생물들의 그것과 같다는 것을 알 수 있다.

앞에서 알아본 몇 가지 조사 결과를 통해, 생물체 유전체 염기서열에서 A 의 비율과 T 의 비율, C 의 비율과 G 의 비율이 서로 비슷하다는 것을 알게 되었다. 또, 생물체 유전체 내에 포함되어 있는 각종 유전자의 염기서열의 경우 A 와 T, C 와 G 의 비율이 서로 같지 않음에도 불구하고 전체를 합해서 보면 A 비율-T 비율, C 비율-G 비율 이 비슷하게 나온다는 것은 A 를 많이 포함하고 있는 염기 서열 부분이 존재하면 비슷한 길이로 T 를 많이 포함하고 있는 부분도 동시에 존재하고, C 나 G 의 경우에도 같은 규칙을 따르는 것이라고 볼 수 있다.

5. 결론

이 논문을 통해 알아 본 생물체 유전체에서의 base-composition 에 대한 정보는 일부 염기서열이 밝혀지지 않은 유전체가 있을 때, 그 부분의 염기서열에 A,C,G,T 가 대략 어떠한 비율로 분포할 것인가를 예측할 수 있는 등 응용 가능성이 매우 크다. 또, 이 정보는 염기서열 분석 프로젝트가 제대로 진행되고 있는가를 대략적으로 알아보기 위한 단서로도 사용될 수 있는데, 앞의 [표 6]과 같은 결과는 *Z.mobilis* 의 유전체에 대한 염기서열 분석 작업이 성공적으로 진행되었음을 뒷받침해 주는 증거의 하나로 볼 수 있다.

6. Acknowledgements

(주)마크로젠에 재직 중이신 ' 강지은' 씨와 ' 이환석' 씨께 감사 드린다.

7. 참고문헌

[1]. Masahiko M, Minoru K, "The Distribution Profiles of GC Content Surrounding the Translation Initiation Site in Different Species", *Proceedings of Genome Informatics Workshop III*, Japan, 1992.

[2]. Boguski M.S, "The turning point in Genome research", *Trend Biol. Sci.* 20, 295-296, 1995.

[3]. Strachan T, Read A.P, "*Human Molecular Genetics*", BIOS Scientific, Oxford, 1996.

[4]. McVean G. A, Hurst G. D, "Evolutionary lability of context-dependent codon bias in bacteria", J Mol Evol, Mar;50(3), 264-275, 2000

[5]. Oliver, J.L., A. Marín. "A relationship between GC content and coding-sequence length", J Mol Evol 43(3): 216-223, 1996.