Nearest-Neighbor Model을 이용한 DNA chip 설계 알고리즘

제갈성준, 문일†
연세대학교 화학공학과

DNA Chip Design Algorithm Using Nearest-Neighbor Model

Sungjoon Chegal and Il Moon†
Department of Chemical Engineering, Yonsei University

서론

DNA chip(또는 DNA microarray)은 각종 생명체가 가지고 있는 방대한 양의 유전정보를 빠른 시간에 대량으로 분석, 처리할 수 있는 신기술이자 현재의 Genomics 열풍을 가능하게 하는 강력한 무기라 할 수 있다[1,2]. 그 중 Gene Expression Array는 유전자 발현분석을 하는데 쓰이는 DNA chip으로서 mRNA에 대한 역전사 기술을 이용해서 만들어진 cDNA를 target DNA로 사용하므로써 유전자의 발현정도를 분석할 수 있는 가장 일반적인 DNA chip이다. 이때 DNA chip에 붙어 있는 probe의 선택은 DNA chip의 설계과정에서 가장 중요한 단계이다. 본 논문에서는 Nearest-Neighbor Model을 이용한 probe의 열역학적 특성을 계산하는 방법과 이를 기반으로 각 cDNA에 대한 최적의 probe를 찾는 설계 알고리즘을 제시하였다.

1. Gene Expression Microarray

일반적으로 DNA chip이라고 일컬어 지는 것은 유리나 그밖의 고정물에 Sequence를 알고 있는 DNA 조각을 붙여서 target DNA와 함께 hybridization 시킴으로써 target DNA의 여러 가지 특성을 분석할 수 있는 실험도구를 의미한다. 그 중에서도 gene expression microarray는 DNA chip의 가장 일반적인 용도인 유전자 발현분석에 쓰이는 것이다. **Figure 1**은 gene expression microarray의 사용방법에 대한 그림이다[3]. 미리 제작된 microarray에는 수천~수만개의 probe가 부착되어 있으며 이들의 염기서열은 알고 있는 상태이다. 여기에 사용되는 target DNA로 어떤 것을 사용하느냐에 따라서 DNA chip의 용도가 달라지는데 gene expression microarray의 경우는 각기 다른 세포에서 발현된 mRNA를 역전사함으로써 얻어진 cDNA를 target DNA로 사용하게 된다. 그림과 같이 tumor cell에서 발현되는 여러종류의 mRNA를 추출하고 또한 정상 세포에서도 발현되는 여러종류의 mRNA를 추출한다. 이들에 각각 역전사 효소를 작용시켜 cDNA를 만들어 내고 여기에 Cy5와 Cy3의 서로 다른 색을 가진 형광물질을 부착시켜 주면 각각의 cDNA는 자신이 발생하게 된 세포를 구분할 수 있는 형광표식을 가지고 있게 된다. 이 target DNA를 DNA chip 위에 뿌려주게 되면 서로 상보적인 관계에 있는 target DNA

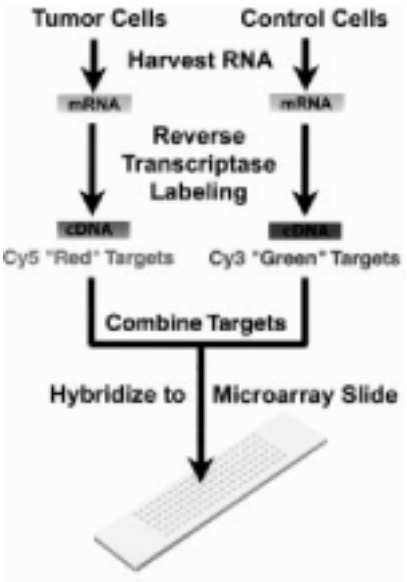


Figure 1. Gene Expression Microarray[3]

와 probe간의 hybridization이 일어나게 되는데 형광색을 관찰할 수 있는 스캐너를 이용하여 각각의 probe spot이 가지고 있는 색을 관찰하므로써 각 세포에서 발현된 mRNA에 대한 유전정보를 얻을 수 있게 된다.

2. Nearest-Neighbor Model

Nearest-Neighbor Model은 DNA hybridization이 일어난 상태에서 각각의 염기서열에 대한 정보로부터 전체 DNA 조각의 열역학 정보를 얻을 수 있는 방법이다. **Figure 2**는 hybridization이 일어난 두 DNA 조각에서의 nearest neighbor을 나타내고 있다[4]. 여기서 각각의 nearest neighbor가 가지고 있는 parameter를 구하기 위해서 용액 내에서의 hybridization 실험을 하게된다. **Figure 3**은 각각의 다른 probe의 농도(C_T)에

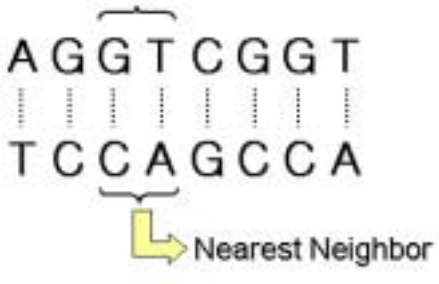
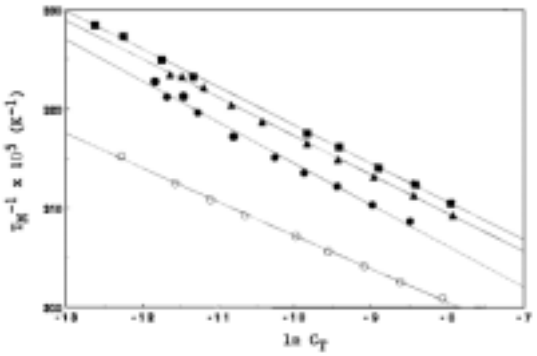


Figure 2. Nearest Neighbor



$$T_M = \Delta H^\circ / (\Delta S^\circ + R \ln C_T)$$

$$T_M^{-1} = R/\Delta H^\circ \ln C_T + \Delta S^\circ / \Delta H^\circ$$

Figure 3. T_m^{-1}과 ln C_T의 관계 [5]

서 측정된 Melting Temperature(T_m)에 대한 실험자료를 plot한 것이다[5]. y축을 T_m^{-1}으로, x축을 ln C_T으로 하여 plot하게 되면 기울기로부터 R/△H를 구할 수 있으며 y절편으로부터 △S/△H를 구할 수 있게 된다. 여러 종류의 sequence를 가진 DNA 조각에 대하여 실험값을 모아 각각의 nearest neighbor 들에 대한 파라미터를 구하게 된다.

3. 알고리즘

반응실험에서 사용하기를 원하는 DNA chip의 반응온도 조건을 입력하고 target DNA에 대한 염기서열 데이터를 입력받아서 순차적으로 계산할 probe를 선택하게 된다. Nearest-neighbor model을 이용하여 T_m을 계산하고 이와 반응온도 조건과의 적합성 유무를 검사한 뒤 적합 판정이 내려지면 probe candidate pool에 포함시킨다. 만약 적합하지 않으면 다른 probe를 선택하여 같은 계산 과정을 거치게 된다. 최종 probe까지 모두 계산과 판정을 거치게 되면 probe candidate pool이 확정되며 다시 Nearest-Neighbor model을 이용하여 pool 내부에 있는 probe들에 대한 Gibbs free energy를 계산하게 된다. 그 결과로 가장 낮은 값을 가진 probe를 입력된 target DNA에 대한 최적 probe로 결정하고 저장하게 된다. 이 알고리즘은 MATLAB™을 이용하여 구현하였으며 내부에는 Nearest-neighbor model parameter[6]가 내장되어 있다. 입력되는 target DNA는 FASTA[7] 포맷의 파일상태로

loading 할 수 있도록 하였다.

4.계산결과

Target DNA로 입력한 DNA sequence는 A-2 plaque virus의 유전자이며 반응온도조건은 40℃로 입력하였을 때의 결과값을 **Table 1**에 나타내었다. 1차 반응온도조건을 통과한 probe들이 probe candidate pool에 모여있게 되는데 이들에 대한 데이터는 **Figure 5**와 **Figure 6**에 나타내었다. 가장 낮은 △G 값을 갖는 probe는 98번째의 probe이며 pool에 선택된 probe들의 T_m은 40~47℃의 범위임을 확인 할 수 있다.

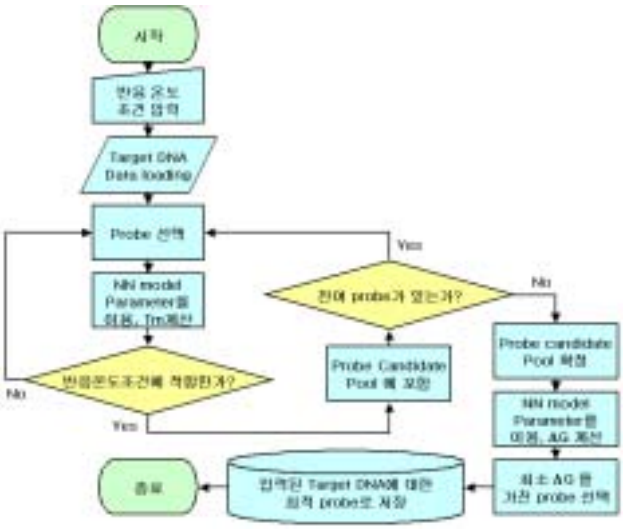


Figure 4. Probe Design Algorithm

입력	Target DNA	>gi|21389216|ref|NC_003988.1| A-2 plaque virus, complete genome
	Sequence	GAGTGTTCCCACCCAACAGGCCCACTGGGTGTTGTACTCTGGTATTACGGTACCTTTGTACGCCTATTTTATTTCCCCCCCCTTTTTGAAACTTAGAAGTTAATAATAAACACGCTCACTAGGTGCACTACATCCAGTAG
	반응온도조건	40℃
결과	최적 probe	TCAATTATTATTTG
	Hybridization 위치	GAGTGTTCCCACCCAACAGGCCCACTGGGTGTTGTACTCTGGTATTACGGTACCTTTGTACGCCTATTTTATTTCCCCCCCCTTTTTGAAACTTAGAAGTTAATAATAAACACGCTCACTAGGTGCACTACATCCAGTAG
	△G	-14.158 Kcal/mol
	Tm	44.4℃

Table 1. 140bp의 DNA sequence 입력에 대하여 계산된 결과값

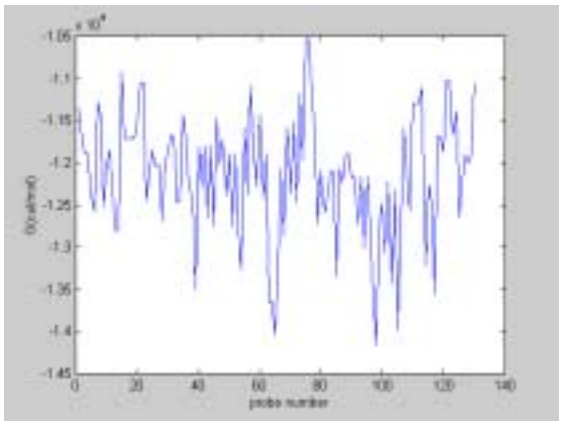


Figure 5. △G values in the pool

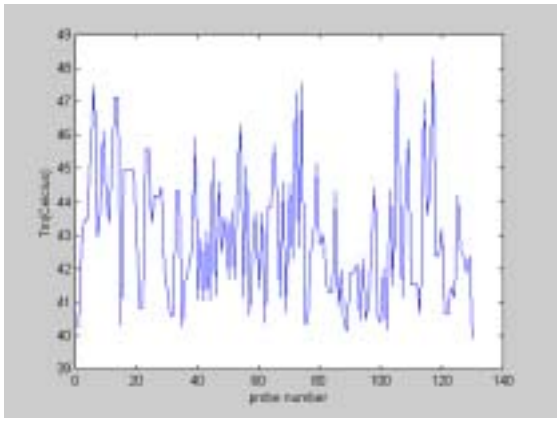


Figure 6. T_m values in the pool

결론

Probe design 과정에서 target DNA에 가장 적합한 probe를 선택하는 것은 매우 중요한 단계이다. 본 연구에서는 T_m과 △G를 이용하여 최적 probe를 선택하는 알고리즘을 개발하였으며 각 probe의 열역학적인 데이터를 얻는 과정에서 Nearest-neighbor model을 사용하였다. 계산된 결과값을 probe candidate pool에 저장함과 동시에 최적의 probe를 찾아낼 수 있었다.

참고문헌

1. 이상엽, 윤성호, 임근배, 조윤경, Genomics와 DNA chip, *News & Information for Chemical Engineers*, **18**, 3, 307-311, 2000
2. S.J. Watson, et al., The ″Chip″ as a Specific Genetic Tool, *Biol. Psychiatry 2000*, **48**, 1147-1156, 2000
3. Javed Khan, et al., Expression profiling in cancer using cDNA microarrays, *Electrophoresis*, **20**, 223-229, 1999
4. Crothers, D. M. and Zimm, *B. H., J. Mol. Biol.*, **9**, 1-9, 1964
5. J. Santalucia et al., *Biochemistry*, **35**, 3555-3562, 1996
6. J. Santalucia, **Proc. Natl. Acad. Sci. USA**, **95**, 1460-1465, 1998
7. National Center for Biotechnology Information(http://www.ncbi.nlm.nih.gov/)