CpG Island 검색용 윈도우 프로그램 개발

김 기 봉*

상명대학교 공과대학 생명정보공학과

Received June 9, 2008 / Accepted August 24, 2008

Development of a Window Program for Searching CpG Island. Ki-Bong Kim*. *Department of Bioinformatics Engineering, Sangmyung University, Chunan 330-720, Korea* - A CpG island is a short stretch of DNA in which the frequency of the CG dinucleotide is higher than other regions. CpG islands are present in the promoters and exonic regions of approximately 30~60% of mammalian genes so they are useful markers for genes in organisms containing 5-methylcytosine in their genomes. Recent evidence supports the notion that the hypermethylation of CpG island, by silencing tumor suppressor genes, plays a major causal role in cancer, which has been described in almost every tumor types. In this respect, CpG island search by computational methods is very helpful for cancer research and computational promoter and gene predictions. I therefore developed a window program (called CpGi) on the basis of CpG island criteria defined by D. Takai and P. A. Jones. The program 'CpGi' was implemented in Visual C++ 6.0 and can determine the locations of CpG islands using diverse parameters (%GC, Obs (CpG)/Exp (CpG), window size, step size, gap value, # of CpG, length) specified by user. The analysis result of CpGi provides a graphical map of CpG islands and G+C% plot, where more detailed information on CpG island can be obtained through pop-up window. Two human contigs, i.e. AP00524 (from chromosome 22) and NT_029490.3 (from chromosome 21), were used to compare the performance of CpGi and two other public programs for the accuracy of search results. The two other programs used in the performance comparison are Emboss-CpGPlot and CpG Island Searcher that are web-based public CpG island search programs. The comparison result showed that CpGi is on a level with or outperforms Emboss-CpGPlot and CpG Island Searcher. Having a simple and easy-to-use user interface, CpGi would be a very useful tool for genome analysis and CpG island research. To obtain a copy of CpGi for academic use only, contact corresponding author.

Key words : CpG island, hypermethylation, gene prediction, Visual C++, G+C%

서 론

포스트게놈 시대에는 유전자들의 기능 및 조절 메커니즘 규명을 비롯하여, 각 생물체의 게놈 프로젝트를 통해서 얻어진 방대한 DNA 정보와 생명체의 정보 등을 이용한 생명현상 연구가 활발히 이뤄지고 있으며, 특히 DNA가 발현된 단백질의 구조와 기능에 대한 연구, 즉 프로테오믹스(proteomics)가 핵심 분야로 대두되고 있다[15]. 게다가, 지난 수 십년 동안 유전자 발현 및 조절 메커니즘에 대한 연구가 유전학을 바탕으로 진행되어 왔으나 최근 유전자가 발현하는데 DNA 서열과 함께 메틸기가 중요한 역할을 한다는 사실이 밝혀지면서 후성학(epigenetics)이 중요한 분야로 인식되고 있다. 후성학에서 가장 중요한 역할을 하는 것은 DNA 염기에 메틸기(CH₃-)가 붙는 메틸화(methylation)이다. 유전체의 염기서열 중 시토신(C)과 구아닌(G) 두 염기가 나란히 존재하는 것을 CpG라 하는데[1], 여기서 시토신이 메틸화될 수 있다[21]. CpG는 진화과정을 거치면서 진핵생물 유전체에서는 점진적

으로 감소되어 온 경향이 있는데 포유동물 유전체에는 여전히 잘 보존되어 있으며 유전체 내 CpG의 메틸화 정도나 양상은 포유동물의 종에 따라 다르고, 조직에 따라서도 특이적인 것으로 밝혀지고 있다[6,16,17,25]. 특히 인간을 포함한 포유동물의 염기서열에는 CpG가 밀집되어 있는 부위가 존재하는데, 이러한 영역을 'CpG island'라고 한다. CpG island는 보통 0.5-4 kbp 정도 크기로, 유전자의 상위 부위, 즉 유전자의 전사를 조절하는 영역인 프로모터 부근에 대부분 위치하고 있어 유전자와 밀접한 연관성을 갖고 있다[2]. 현재 밝혀진 인간 유전체의 전체 염기서열에서 CpG island는 약 29,000개로 추정되는데, 전체 추정 유전자의 70%에 해당하는 유전자가 하나의 CpG island를 보유하고 있는 것으로 여겨진다. 많은 실험들에 의해 CpG island는 유전자 발현억제[3], 유전자 각인[8,26], X 염색체 불활성화[19,23], 발암의 억제[14] 등에 작용 하는 등 유전자 발현 및 조절 메커니즘에 매우 중요한 역할을 하는 것으로 밝혀졌다.

CpG island는 대부분 비메틸화상태(methylation-free)로 존재한다[2,20]. 그러나 메틸화된 예외적인 경우가 존재해 CpG island의 메틸화가 지닌 의미를 밝히는데 도움이 되고 있다. 특히, 인간의 정상세포에서는 암 억제유전자 등이 정상

*Corresponding author

Tel : +82-41-550-5377, Fax : +82-41-550-5184

E-mail : kbkim@smu.ac.kr

적으로 발현됨으로써 세포분열을 적절하게 조절하고 있다. 그러나 암세포의 경우는 암 억제유전자가 제대로 기능을 발휘하지 못해 세포분열의 조절기능을 잃어버린 상태에 이른다[9,11]. 이런 암 억제유전자의 기능 소실은 이들 유전자의 CpG island에 나타난 비정상적 메틸화에 원인이 있다는 사실이 다양한 암(대장암[10], 간 중앙[24], 백혈병[13], 위암[5,22] 등) 중에서 연속적으로 밝혀지고 있다[18]. 유전자의 전사과정은 세 가지 조건이 충족되어야 진행이 되는데 첫째, 적절한 전사인자(transcriptional factor)가 있어야 하고, 둘째, 히스톤(histone)들이 아세틸화되어 있으면서 메틸화되지 않아야 하고, 셋째, CpG island에 메틸화가 이뤄지지 않아야 한다[7]. 이러한 측면에서 CpG island가 매우 중요한데, 비록 CpG island의 메틸화가 우선인지 아니면 유전자 발현 억제가 우선인지의 논란은 여전히 남아 있지만[4], 암세포에서 CpG island내의 메틸화는 분자진단 및 암의 조기 진단에 활용될 가능성이 매우 높으며, 유전자의 변이보다는 후생학적 변이를 정상화시키는 것이 더 쉬운 것이라는 점에서 암의 치료적 측면에서도 매우 흥미로운 분야이다. 따라서 이런 다양한 연구결과들이 밝혀짐에 따라 암과 관련하여 CpG island를 연구하는 곳이 늘어나고 있는 실정이다. 게다가, CpG island는 유전자의 유용한 표식자로서의 역할을 할 수 있기 때문에 전산생물학적 측면에서 유전자나 프로모터 예측 프로그램을 개발할 때 중요한 단서로 활용될 수 있다.

전산학적 기법을 이용한 CpG island 검색 프로그램은 대규모 유전체 프로젝트의 산물로 쏟아지고 있는 유전체를 대용량으로 분석할 때 유용한 분석 도구로 활용될 수 있을 뿐만 아니라, CpG island 관련 연구자들에게도 많은 도움이 될 수 있을 것이다. 대표적으로 널리 활용되고 있는 공개용 웹 기반 검색 프로그램으로는 D. Takai와 P.A. Jones [12,27]에 의해서 구현된 CpG Island Searcher (<http://cpgislands.usc.edu/>)와 Gardiner-Garden와 Frommer가 제안한 기준[12]으로 구현된 EMBOSS의 CpGPlot (<http://www.ebi.ac.uk/emboss/cpgplot/#andNewcpgseek>) 등이 있다. 국내에는 아직까지 공개된 웹 프로그램이 존재하지 않으며, 게다가 국내외적으로 윈도우 운영체제의 개인용 컴퓨터 상에서 독자적으로 구동될 수 있는 윈도우 기반의 검색 프로그램이 공개되어 있지 않은 실정이다. 이러한 측면에서 쉽게 사용가능하며 분석결과에 대한 해석이 용이한 사용자 친화적인 인터페이스 기반의 CpG island 검색용 윈도우 프로그램인 CpGi를 개발하게 되었다.

재료 및 방법

개발 내용 및 개발 환경

CpGi는 유전체를 대용량으로 분석하고 주석 작업을 수행하는 연구자들이나, CpG island의 메틸화를 바탕으로 프로모터나 각종 암 관련 유전자의 발현 유무를 연구하는 실험 연구

자들을 위하여 보다 용이하고 정확하게 CpG island를 예측할 수 있도록 개발된 윈도우 프로그램이다. 현재 일반에 공개되어 널리 사용되고 있는 CpG island 검색용 웹 프로그램들의 경우 텍스트 위주의 분석 환경을 제공하고 있으며, 비록 부분적인 그래픽 환경을 제공하고 있더라도 인터넷 웹상의 한계로 말미암아 분석결과를 효과적으로 가시화(visualization)하지 못하는 실정이다. 따라서 이러한 문제점을 극복하고 다양한 파라미터 설정에 의한 효과적인 분석과 분석결과에의 가시화를 위해 최대한 사용자 친화적인 윈도우 인터페이스(Interface)로 CpGi 프로그램을 구현하였으며, 윈도우 프로그램이기에 마이크로소프트 윈도우즈 운영체제의 개별 PC 상에서 독자적인 프로그램으로 구동된다. 프로그램 내부에 사용된 CpG island 검색 기준은 기본적으로는 D. Takai와 P. A. Jones 등이 제안한 기준[12]을 바탕으로 하였고, 이 외에도 부가적인 기준을 첨가하여 사용자가 맞춤형 분석을 할 수 있게끔 다양한 파라미터 값들(%GC, Obs (CpG)/Exp (CpG), 윈도우 크기, 스텝크기, Gap 값, #CpG 등)을 사용자가 직접 설정할 수 있도록 구현하였다(Fig. 1). 윈도우 크기(window size)는 분석대상 입력서열을 분석할 때 매회마다 분석대상으로 삼는 부분 서열(즉, 분석 윈도우)의 크기를 의미하며, 여기서는 CpG island의 최소 크기 값과 동일하다. %GC는 분석 윈도우 내의 염기서열들 중에서 G와 C가 차지하는 백분율 값이며, Obs (CpG)/Exp (CpG)는 임의의 상황에서 이론적으로 추정되는 CpG의 개수에 대한 실제 관측된 CpG의 개수의 비율을 의미한다. 스텝크기(step size)는 선행 분석 윈도우에 대한 분석이 완료되고 그 다음 분석 윈도우로

Fig. 1. Parameters setting window of the CpGi program. Using this window, users can set various parameters for CpG island and G+C% plotting. The parameters for searching CpG island consist of window size, step size, gap size, #CG, G+C% and Obs (CpG)/Exp (CpG).

이동할 때 건너뛰게 되는 서열의 크기를 의미하며, Gap 값은 입력서열 전체에 대한 분석이 완료된 후 검색된 CpG island들에 대해 내부적으로 인접한 CpG island들 간의 단일체로의 합체를 위해 사용되는 합체 허용 서열 간격을 의미한다. 마지막으로 #CpG는 분석 윈도우 내에 존재하는 CpG dinucleotide의 최소 개수 값을 의미한다. 게다가, CpGi는 위치별 G+C%를 그래프로 보여주는 내부 모듈을 갖고 있으며, 이를 위한 별도의 윈도우 크기와 스텝크기 파라미터 값이 존재한다. 즉, CpG island의 검색결과를 시각적으로 보여줄 때 입력서열 상의 CpG island들의 위치와 더불어 입력 서열상의 위치별 G+C%를 그래픽하게 보여주기 위한 옵션이다(Fig. 1). 분석대상 입력서열을 편집박스(edit box)에 직접 입력하거나 파일 형태로 업로드시킬 수 있으며 입력서열은 반드시 FASTA 양식(즉, 첫 번째 줄은 반드시 ' >' 시작하는 주석문이어야 하고, 두 번째 줄부터 실제 서열데이터가 뒤따르는 양식을 의미함)을 원칙으로 하며, 파일의 확장자는 디폴트로 .seq로 설정되어 있다(Fig. 1).

CpGi를 개발하기 위해 Microsoft Windows XP Professional (5.1, build 2600) 운영체제를 탑재한 Intel Pentium 4 (2.80 GHz CPU, 512 MB RAM) 컴퓨터를 사용하였으며, CpGi의 외부 인터페이스와 주요 핵심 알고리즘은 프로그래밍 언어 Visual C++ 6.0을 이용하여 구현하였다.

내부 알고리즘 구현

현재까지 기존의 검색 프로그램에 의해 널리 통용되는 보편적인 CpG island 검색 기준은 크게 두 가지로 대별된다. 즉, Gardiner-Garden과 Frommer에 의해 제시되었던 기준[12]과 Takai와 Jones에 의해 제안된 기준[27]이다. 전자의 경우는 Gardiner-Garden과 Frommer가 GenBank 서열 데이터를 대량으로 분석한 것을 토대로 1987년에 제시한 기준으로 지금까지 일반적으로 널리 통용되고 있다. 그들이 제시한 기준에 의하면 CpG island는 기본적으로 서열의 길이가 적어도 200 bp 이상이어야 하며 해당 길이의 서열 내에서 G+C%가 50% 이상이고, 추정되는 GpG 수에 대한 관측된 GpG의 비율(즉, Obs (CpG)/Exp (CpG))이 0.6 이상이어야 한다. Takai와 Jones에 의해 2002년에 제시된 기준을 살펴보면 CpG island는 기본적으로 서열의 길이가 500 bp 이상, 서열 내의 G+C%는 55% 이상, 그리고 Obs (CpG)/Exp (CpG)의 비율이 0.65 이상이어야 한다. 따라서 Takai와 Jones에 의해 제시된 기준은 Gardiner-Garden과 Frommer에 의해 제시된 기준을 다소 보완하고 개선하여 검색 대상 서열내의 CpG island를 보다 엄격한 기준으로 정확하게 찾아낼 수 있도록 제안된 것이라고 볼 수 있다. 이러한 측면을 고려하여 본 논문에서는 Takai와 Jones가 제시한 기준을 토대로 CpG island 예측 프로그램의 내부 검색 알고리즘을 개발하였다. CpG island를 예측하기 위한 검색 알고리즘은 C 언어를 이용하여

개발하였는데 이를 요약하면 다음과 같다(Fig. 2).

(A) 분석대상 입력서열의 첫 번째 위치를 시작점으로 사용자가 지정한 윈도우 크기만큼 초기 분석 윈도우를 설정한다. 윈도우 크기는 파라메타 설정 화면 창에서 사용자가 임의로 정해줄 수 있으며, 디폴트 값은 200 bp이다.

(B) 설정된 분석 윈도우 내에 위치하는 해당부분 서열들에 대해 %GC, Obs (CpG)/Exp (CpG) 및 #CpG 값들을 계산한다.

(C-1) 만약 분석 윈도우가 CpG island로 판단하는 기준을 충족시키면 분석 윈도우 크기만큼 진행 방향으로 이동시키고, 새로운 분석 윈도우에 대해 (B)에서 계산한 값들을 다시 계산하고 해당 영역이 기준을 충족시키는지 여부를 판단한다. 기준이 충족되어질 경우 반복적으로 분석 윈도우 크기만큼 진행 방향으로 이동하고 새로운 분석 윈도우에 대해 계산하는 반복 작업을 수행하면서 CpG Island 영역을 확장해 나간다. 이러한 반복적 확장과정에서 기준에 부합되지 않는 분석 윈도우가 발생하면 진행방향의 역방향으로 스텝크기 만큼씩 분석 윈도우를 이동시키면서 기준이 부합되는 영역이 나올 때 까지 계속 역방향으로 진행한다.

(C-2) 단계 (B)에서 계산한 결과 값들이 CpG island로 판단하는 기준을 충족시키지 못 할 경우 분석 윈도우를 스텝크기 만큼씩 진행 방향으로 이동시키면서 기준을 충족시킬 때

Fig. 2. Schematics for the algorithms for searching CpG islands. The main procedure can be described as follows: (A) Set a 200-base window in the beginning of an input sequence, compute each parameter value. Shift the window 1 bp after evaluation until the window meets the criteria of a CpG island. (B) If the window meets the criteria, shift the window 200 bp and then evaluate again. (C) Repeat these 200-bp shifts until the window does not meet the criteria. (D) Shift the last window 1 bp toward the 5' end until it meets the criteria. (E) If this large CpG island does not meet the criteria, trim this CpG island 1 bp from 5' side until it meets the criteria. (F) Merge two individual CpG islands if they are separated by less than specified gap value.

까지 계속 진행한다. 충족되어지는 영역이 검색되어지면 (C-1)단계의 작업을 수행한다.

(D) 분석 윈도우의 끝부분이 입력 서열의 마지막 위치에 도달하면 (E)단계로 이동한다. 그렇지 않으면 분석 윈도우의 끝부분에 새로운 분석 윈도우를 위치시키고 (B)단계에서 (D) 단계 과정을 반복 수행한다.

(E) 검색 완료된 입력 서열내의 각각의 CpG island에 대해 인접한 것들끼리의 거리가 Gap 값 미만이고 상호 연결했을 경우에 합쳐진 단일 영역이 %GC 및 Obs (CpG)/Exp (CpG) 값들을 충족시키는 것들이 있으면 서로 연결하여 하나의 단일 CpG island로 최종 보고한다.

CpG Island Searcher와 Emboss-CpGPlot 프로그램들의 검색 성능 비교

CpGi의 검색 성능을 파악하기 위해 기존의 검증된 웹 인터페이스 기반의 CpG island 검색 프로그램인 CpG Island Searcher와 Emboss-CpGPlot 등과 동일하거나 거의 유사한 파라미터 내에서 검색 결과를 비교분석 하였다. 실제 각 프로그램 마다 사용할 수 있는 파라미터와 내부 검색 알고리즘 등의 미묘한 차이 등으로 말미암아 100% 완전한 비교분석을 할 수 없지만 나름대로 객관성을 확보하기 위해 파라미터들을 최대한 유사하게 설정하고 비교분석하였다. 각 프로그램의 파라미터들을 살펴보면, 먼저 Takai와 Jones의 기준에 기반을 둔 CpG Island Searcher의 경우 모든 파라미터들은 라디오 버튼 형태로 구성되어 있다. %GC는 50%부터 시작해서 60%까지 1% 간격으로 11개의 라디오 버튼으로 이뤄져 있으며 기본 값으로는 55%로 설정되어 있고, Obs (CpG)/Exp (CpG)는 0.6부터 0.7까지 0.01 간격으로 11개의 라디오 버튼으로 구성되어 있으며 기본 값으로는 0.65로 설정되어 있다. 그리고 분석 윈도우 크기는 200 bp부터 1,200 bp까지 100 bp 간격으로 선택할 수 있으며 기본 값으로는 500 bp로 설정되어 있고, 마지막으로 Gap값은 100 bp부터 300 bp까지 50 bp 간격으로 설정할 수 있으며 기본 값으로는 100 bp를 사용하고 있다. Emboss-CpGPlot은 기본적으로 Gardiner-Garden와 Frommer의 기준에 기반을 두고 있으며 분석 윈도우 크기는 100 bp 부터 400 bp 까지 100 bp 씩, Obs (CpG)/Exp (CpG)는 0.6부터 1.0 까지 0.1씩, MinPC는 30부터 80까지 10씩 (MinPC는 CpG island로 확정짓기 전에 10개의 분석 윈도우 내에서 G와 C가 차지하는 최소 평균 백분율 값을 의미함), CpG island의 최소 길이 값인 Length는 50 bp, 100 bp, 200 bp, 300 bp, 400 bp, 500 bp 등의 값을 선택할 수 있으며 그 외에 스텝크기로는 1, 2, 5, 10 등의 값을 선택할 수 있게 되어 있다. 이렇듯 앞에서 언급한 바대로 CpG island를 검색하기 위해 개발된 프로그램들의 특성과 내부 알고리즘의 기준에 따라 각각의 파라미터들이 다르기 때문에 CpGi가 유효한 결과를 나타내는지 비교분석하기 위해 최대한 유사하거나

동일한 파라미터 값들을 적용하도록 하였다. 그러한 측면에서 Emboss-CpGPlot의 일부 파라미터들은 선택할 수 있는 값들의 차이가 너무 크기 때문에 Emboss-CpGPlot의 기준에 CpG Island Searcher와 CpGi의 기준을 최대한 맞추어 결과를 분석하였다. 비교 분석 시 22번 염색체의 AP00524 Contig와 21번 염색체의 NT_029490.3 Contig를 사용하였고, 설정한 파라미터 값들은 Emboss-CpGPlot의 경우 분석 윈도우 크기는 200 bp, 스텝크기는 1 bp, Obs (CpG)/Exp (CpG)는 0.6, MinPC는 50, Length는 200 bp였고, CpG Island Searcher의 경우에는 %GC는 55, Obs (CpG)/Exp (CpG)는 0.6, Length는 200 bp, Gap값은 100 bp였다. 마지막으로 CpGi의 경우 분석 윈도우 크기는 200 bp, 스텝크기는 1 bp, Obs (CpG)/Exp (CpG)는 0.6, Gap값은 100 bp, %GC는 55, 그리고 #CpG는 10으로 설정하여 실행하였다.

결 과

CpGi 검색 결과의 양식 및 해석 방법

CpGi 검색 결과의 양식과 해석 방법에 대한 이해를 돕기 위해 13번 염색체의 NC_006595 (1~3846) Contig를 사용하여 CpGi를 실행하였다. 검색 결과 화면을 살펴보면 위쪽 라인은 분석대상 입력서열을 의미하고, 검색된 CpG Island은 굵은 청색선으로 표시된다. 그리고 분석대상 입력서열 라인 바로 밑 부분에는 해당 입력서열의 위치별 G+C% 값들을 그래픽하게 도표화해서 보여준다. G+C% 도표는 사용자가 다양한 용도로 참조하거나 활용할 수 있을 것이다. CpG island의 굵은 청색선 부분을 더블클릭하면 해당 CpG island에 대한 세부정보를 보여주는 팝업 창이 뜬다. 즉, 팝업 창을 통해서 CpG island의 시작위치와 종료위치, CpG island의 전체 길이, G+C%값 및 CpG island를 구성하고 있는 실제 서열 등을 확인할 수 있다(Fig. 3). Fig. 3을 통해서 알 수 있듯이 NC_006595에서는 모두 세 개의 CpG island가 검색되었다.

22번 염색체의 AP00524 Contig에 대한 검색 결과

22번 염색체의 AP00524 (1~4455) Contig를 대상으로 Emboss-CpGPlot, CpG Island Searcher 및 CpGi 등을 구동시켰을 경우, 모든 프로그램들이 Contig 내에 CpG island가 존재하지 않는 것으로 검색결과를 보고하였다(Fig. 4).

21번 염색체의 NT_029490.3 Contig에 대한 검색 결과

21번 염색체의 NT_029490.3 (1~15050) Contig를 대상으로 세 개의 프로그램을 구동시켰을 경우, Emboss-CpGPlot는 하나의 CpG island가 존재하는 것으로, CpG Island Searcher와 CpGi는 각각 2개의 CpG island가 존재하는 것으로 검색결과를 보고하였다. Emboss-CpGPlot에서 찾아낸 CpG island의 길이는 225 bp이었고, 시작 위치는 입력서열

Fig. 3. (A) Graphical window interface of CpGi for displaying analysis result. The contig NC_006595 derived from human chromosome 13 was analyzed to show the format of analysis result. Three CpG islands were detected to be displayed in thick line bar and the positional G+C% has been plotted below input sequence line. (B) Pop-up window of CpGi to show more detailed information on CpG island. User can double-click over a CpG island on the result screen to open this pop-up window which provides start point, end point, length, G+C% and sequence data of the CpG island.

의 13555번째이고 종료 위치는 13779번째이었다. CpG Island Searcher가 검색한 첫 번째 CpG island는 입력서열의 10287번째에서 10440번째 사이에 위치했으며, %GC는 55.2%, Obs (CpG)/Exp (CpG)는 0.607, 길이는 154 bp이었다. CpG Island Searcher가 검색한 두 번째 CpG island의 위치는 13575에서 시작되어 13890에서 끝났다. %GC는 55.1%,

Fig. 4. Analysis results of Emboss-CpGPlot, CpG Island Searcher and CpGi on the contig AP00524 from chromosome 22. (A), (B), and (C) represent the analysis results of Emboss-CpGPlot, CpG Island Searcher and CpGi respectively. They all reported that there is no CpG island on the contig AP00524.

Obs (CpG)/Exp (CpG)는 0.851 그리고 길이는 316 bp이었다. 마지막으로 CpGi의 검색 결과를 살펴보면 검색된 첫 번째 CpG island는 10218에서 시작해 10437에서 끝났으며 그 길이는 219 bp, %GC는 51%를 나타냈다. 검색된 두 번째 CpG island는 13464에서 시작해 13869에서 끝났으며 길이는 405 bp, %GC는 48%였다(Fig. 5).

Fig. 5. Analysis results of Emboss-CpGPlot, CpG Island Searcher and CpGi on the contig AP00524 from chromosome 22. (A), (B), and (C) represent the analysis results of Emboss-CpGPlot, CpG Island Searcher and CpGi respectively. Even though Emboss-CpGPlot reported that there exists one CpG island on this contig, CpG Island Searcher and CpGi stated that there exist two CpG islands.

고 찰

첫 번째 검증 실험에 해당하는 22번 염색체의 AP00524 Contig에 대한 검색결과를 살펴보면 CpGi가 CpG Island Searcher와 Emboss-CpGPlot 등과 일치하는 검색결과로서 해당 Contig상에 하나의 CpG island도 존재하지 않는다는 검색결과를 도출하였다. 원래 22번 염색체의 AP00524는 CpG island가 존재하지 않은 Contig였기 때문에 False positive한 것들을 예측해 내지 않고 나름대로 정확하게 예측했다고 할 수 있다. 즉, 특이성(specificity) 측면에서 나름대로 정확성을 띄고 있다는 것을 간접적으로 확인할 수 있었다. 그러나 이러한 하나의 결과를 가지고 CpGi가 기존의 다른 프로그램들만큼 성능이 뛰어나다고 단정하기 힘들 것이다. 특히 False negative한 예측치 측면에서의 정확성, 즉 민감도(sensitivity) 차원에서 다른 두 프로그램들 보다 뛰어나다고 단언하기 곤란할 것이다. 이미 앞에서 언급한 바가 있듯이, 실제 CpG island 검색 프로그램들은 나름대로의 기준에 의해서 자체 고유의 내부 알고리즘으로 검색하기 때문에 상호 우위를 직접적으로 비교하기 매우 힘들며, 특정한 기준을 정하여 검색했을 때 합당한 것들을 모두 검색해 낼 수 있는 것이 보편적이기 때문에 더욱 그러하다. 그래서 그러한 성능평가는 다른 응용 분야의 프로그램들에 비해 그리 중요하지 않다고 볼 수 있다. 그럼에도 불구하고, False negative한 예측치 측면의 성능에 대해 간접적으로나마 검증하기 위해 21번 염색체의 NT_029490.3을 이용하여 검색한 결과들을 비교분석해 보았다. 21번 염색체의 NT_029490.3에 대해서는 Emboss-CpGPlot는 하나의 CpG island를 예측하였고, CpGi와 CpG Island Searcher는 두 개의 CpG island가 존재하는 것으로 예측하였다. 3개의 분석 프로그램들이 예측한 CpG island의 개수에 차이가 있을 뿐만 아니라 예측된 CpG island들의 위치가 검색 프로그램별로 조금씩 상이했다. 프로그램마다 이러한 상이한 위치의 차이가 생기는 것은 아마도 파라미터의 지정문제와 내부 검색 알고리즘의 미묘한 차이에서 기인한다고 볼 수 있을 것이다. 즉, 각 프로그램별로 기본적으로 주어진 파라미터의 종류에 다소 차이가 있고, 동일한 파라미터라고 하더라도 지정할 수 있는 수치의 단위나 간격의 차이가 있었기 때문에 이러한 결과가 나오게 된 것이라 여겨진다. 그렇지만 이러한 결과가 유의하다고 생각되는 것은 CpGi를 통하여 검색된 CpG island의 위치가 동일하게 두 개의 CpG island를 찾아낸 CpG Island Searcher의 결과와 비교해 봤을 때 거의 비슷한 곳에 위치하고 있었고, 한 개의 CpG island를 검색해 낸 Emboss-CpGPlot의 분석 결과의 도표를 잘 분석해 보면 CpG Island Searcher와 CpGi에서 검색된 첫 번째 CpG island의 Obs (CpG)/Exp (CpG) 및 GC% 등의 파라미터 값들과 유사한 값을 갖고 있다는 것을 발견할 수 있었다. 다른 두 프로그램이 검색한 하나의 CpG island를 Emboss-CpGPlot가 제대로 검

색하지 못했던 것은 앞서 언급한 바와 같이 사용자가 지정할 수 없는 분석 시스템 내부 파라미터의 적용 방식에 다소 차이가 있었기 때문에 CpG Island Searcher나 CpGi에서는 유효하다고 여겨진 CpG island 영역이 Emboss-CpGPlot에서는 유효하다고 판단되는 임계값(threshold)에 도달할 수 없었던 것으로 이해할 수 있다.

서론에서 언급한 것처럼 CpG island 메틸화는 분자 진단 및 암의 조기 진단에 활용될 가능성이 매우 높으며, 유전자의 변이보다는 후성학적 변이를 정상화시키는 것이 더 쉬울 것이라는 점에서 암의 치료적인 측면에서도 응용 가능성이 높다고 할 수 있다. 특히 메틸화 여부 혹은 그 양상의 차이 등은 같은 암을 가진 환자들에게서 예후를 짐작할 수 있게 하거나 항암제에 대한 반응 등을 예측하는데 사용될 수도 있을 것이다. 즉, 현재로써는 수많은 CpG island와 그와 관련되어 있는 메틸화 중에서 임상적으로 유용한 것들을 탐색하는 단계에 있지만 머지않은 장래에는 DNA 메틸화 표지가 암 위험도 측정, 조기 진단, 암 조직에 대한 분자 진단, 항암 치료 효과 예측, 재발 감시 등에 광범위하게 응용될 수 있을 것이다. 전산학적 측면에서 볼 때 대규모 유전체 프로젝트로 말미암아 쏟아지는 DNA 서열 데이터에 대한 기본적인 분석 및 주석 작업이 필수적이며, 이러한 작업에는 기본적으로 프로모터 예측 및 유전자 예측이 필수적으로 포함된다. 대부분의 CpG island들은 유전자 발현을 제어하는 프로모터 부위에 위치하고 있기 때문에 CpG island를 정확히 검색해 낼 수 있다면, 프로모터의 위치나 유전자의 시작지점 등에 대한 중요한 단서를 찾을 수 있을 것이다. 이러한 맥락에서 순수한 CpG island 연구뿐만 아니라 대규모 서열 데이터 분석 차원에서 정확한 CpG island를 검색할 수 있는 프로그램이 요구되고 있는 실정이다. 이에 따라 서버와 클라이언트 환경의 웹 기반 프로그램이 아닌 윈도우 운영체제의 개인용 컴퓨터 상에서 구동될 수 있는 CpG island 검색용 윈도우 프로그램을 개발하게 되었다. 프로그램의 개발 도구로는 Visual C++ 6.0을 사용하였고, 검색 알고리즘은 Takai와 Jones가 제시한 기준을 기반으로 개발되었다. CpGi의 검색 결과에 대한 객관적 판단을 위해 22번 염색체의 AP00524와 21번 염색체의 NT_029490.3를 이용하여 CpG Island Searcher와 Emboss-CpGPlot에서 도출된 분석결과와 비교 분석해 보았다. 앞에서 보았듯이, CpGi에 의한 검색 결과는 다른 두 프로그램에 비해 같거나 오히려 보다 정확한 검색 결과를 보여주었다. Emboss-CpGPlot와 CpG Island Searcher의 대부분의 파라미터 값들은 사용자들이 이미 고정적으로 제공된 라디오버튼 형태의 파라미터 값을 선택하는 방식으로 되어 있으나, CpGi의 경우 기본 값들은 Takai와 Jones의 기준에 의해 지정되어 있으면서도 편집박스 형태로 사용자가 임의의 값을 마음대로 지정해 줄 수 있어 보다 정교한 맞춤형 분석이 가능하다는 것이 큰 장점 중의 하나이다. 게다가, 기존 웹 기반

분석 프로그램들의 취약점인 그래픽 인터페이스의 강화로 사용자들이 보다 쉽게 이용할 수 있고, 분석 결과에 대한 이해를 보다 용이하게 할 수 있다는 것도 중요한 의의로 부각될 수 있다.

요 약

CpG island는 유전자의 발현 기작과 많은 관련이 있다. 포유동물 유전자의 약 30~60% 정도의 프로모터와 엑손부위에 CpG island가 존재 한다. 최근의 연구 결과에 따르면 CpG island의 과메틸화는 주요 암 억제유전자들을 불활성화시켜 암을 일으키는 주요 요인이 되는 것으로 밝혀졌다. CpG island의 과메틸화는 거의 모든 암 종류에서 발견되고 있다. 따라서 CpG island를 검색하는 프로그램은 매우 중요한 의미를 갖는다. 그래서 D. Takai 와 P. A. Jones 등이 2002년 검증한 CpG island 정의 기준을 이용하여 윈도우 기반의 소프트웨어 프로그램인 CpGi를 개발하였다. CpGi는 Visual C++ 6.0로 구현하였으며, 입력서열 양식은 FASTA 포맷을 허용하도록 구성하였다. CpGi의 검색 성능을 평가하기 위해 2개의 인간 Contig, 즉, AP00524 (22번 염색체)와 NT_029490.3 (21번 염색체) 등을 대상으로 기존의 다른 CpG island 검색 프로그램인 Emboss-CpGPlot 및 CpG Island Searcher 등과 검색결과를 비교 분석하였다. CpGi에 의한 검색 결과는 다른 두 프로그램에 비해 같거나 오히려 보다 정확한 검색 결과를 보여주었다. CpGi는 사용자 친화적인 윈도우 인터페이스로 구현되어 있어 사용자가 프로그램을 구동하고 이용하기 매우 쉽고, 분석결과에 대한 이해도 용이하다. 본 프로그램은 사용자가 지정한 파라미터 값들(%GC, Obs (CpG)/Exp (CpG), 분석 윈도우 크기, 스텝크기, Gap 허용치, #CG)에 의해 CpG island의 위치를 결정하고, G+C%와 CpG island의 위치를 시각적으로 보여준다. 결과적으로, CpGi는 CpG island 관련 실험 연구자들뿐만 아니라 대응량 서열 분석 및 주석 작업을 위해 매우 유용한 도구로 활용될 수 있을 것이다.

References

- Antequera, F. and A. Bird. 1993. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* **90**, 11995-11999.
- Bird, A. 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209-213.
- Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes & Dev.* **16**, 6-21.
- Clark, S. J. and J. Melki. 2002. DNA methylation and gene silencing in cancer: which is the guilty party. *Oncogene* **21**, 5380-5387.
- Cravo, M., R. Pinto, P. Fidalgo, P. Chaves, L. Gloria, C. Nobre-Leitao and F. Mira. 1996. Global DNA hypomethylation occurs in the early stages of intestinal type gastric carcinoma. *Gut* **39**, 434-438.
- Ehrlich, M. and R. Wang. 1984. RYH. 5-methylcytosine in eukaryotic DNA. *Science* **212**, 1350-1357.
- Esteller, M. 2003. Relevance of DNA methylation in the management of cancer. *The Lancet Oncology* **4**, 351-358.
- Feil, R. and S. Khosla. 1999. Genomic imprinting in mammals: an interplay between chromatin and DNA methylation. *Trends in Genetics* **15**, 431-435.
- Feinberg, A. P. and B. Volgestein. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**, 89-92.
- Feinberg, A., C. Gehrke, K. Kuo and M. Ehrlich. 1988. Reduced genomic 5-methylcytosine content in human colonic neoplasia. *Cancer Res.* **48**, 1159-1161.
- Gama-Sosa, M., V. Slagel, R. Trewyn, R. Oxenhandler, K. Kuo, C. Gehrke and M. Ehrlich. 1983. The 5-methylcytosine content of DNA from human tumors. *Nucleic Acid Res.* **11**, 6883-6894.
- Gardiner-Garden, M. and M. Frommer. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261-282.
- Hanada, M., D. Delia, A. Alello, E. Stadtmauer and J. C. Reed. 1993. Bcl-2 gene hypomethylation and high-level expression in B-cell chronic lymphocytic leukemia. *Blood* **82**, 1820-1828.
- Jones, P. and S. Baylin. 2002. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3**, 415-428.
- Jung, M., W. Park and K. Kim. 2004. Development of Web-Based Assistant System for Protein-Protein Interaction and Function Analysis. *Journal of Life Science* **14**, 997-1002.
- Kim, K. B. 2007. A Study on the Application Methods of a Support Vector Machine for Gene Promoter Prediction. *Journal of Life Science* **17**, 714-718.
- Kundu, T. K. and M. R. Rao. 1999. CpG islands in chromatin organization and gene expression. *J. Biochem (Tokyo)* **125**, 217-222.
- Laird, P. W. and R. Jaenisch. 1994. DNA methylation and cancer. *Hum. Mol. Genet.* **3**, 1487-1495.
- Latham, K. E. 1996. X chromosome imprinting and inactivation in the early mammalian embryo. *Trends Genet.* **12**, 134-138.
- Li, E., T. H. Bestor and R. Jaenisch. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915-926.
- Nephew, K. and T. Huang. 2003. Epigenetic gene silencing in cancer initiation and progression. *Cancer Letters* **190**, 125-133.
- Ohtani-Fujita, N., T. Fujita, A. Aoike, N. E. Osifchin, P. D. Robbins and T. Sakai. 1993. CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene. *Oncogene* **8**, 1063-1067.
- Panning, B. and R. Jaenisch. 1998. RNA and the epigenetic regulation of X chromosome inactivation. *Cell* **93**, 305-308.
- Rao, P. M., A. Antony, S. Rajalakshmi and D. S. Sarma.

1989. Studies on hypomethylation of liver DNA during early stages of chemical carcinogenesis in rat liver. *Carcinogenesis* **10**, 933-937.
25. Razin, A. and A. D. Riggs. 1980. DNA methylation and gene function. *Science* **210**, 604-610.
26. Razin, A. and R. Shemer. 1995. DNA methylation in early development. *Hum. Mol. Genet.* **4**, 1751-1755.
27. Takai, D. and P. A. Jones. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* **99**, 3740-3745.