보도자료

서울대학교 SEOUL NATIONAL UNIVERSITY

보도일시	즉시 보도
	2024. 7. 25.(목)
문의	담당자: 보건대학원 이은주 (02-880-2741)
	연구책임자 보건대학원 성주헌 교수 (02-880-2828) / 교신저자
	연구진 박건형 연구원 (02-880-2787) / 제1저자

■ 제목/부제

제목	*국문*	*한국인이 개발한 혁신적인 유전체 분석 알고리즘*
부제	*국문*	*유전체 분석의 복잡성과 불투명성을 해결하고 성능과 효율성을 극대화한 새로운 알고리즘*

■ 요약

연구 필요성	유전체 분석술은 질병 진단, 신약 개발 등 바이오테크놀로지의 기반기술이다. 인간게놈프로젝트 당시 13년이 걸렸던 전체염기서열 분석을 현재 하루만에 처리할 수 있게 된 것은 실험분석기술과 함께 자료처리 알고리즘의 기여가 절대적이다. 하지만, 기존의 알고리즘들은 성능확보를 위해 설정값의 복잡성, 알고리즘간의 비호환성이 심화되어, 분석결과의 불투명성과 신뢰성의의 문제가 있었다.
연구성과/ 기대효과	서울대학교 연구진은 생물학적인 의미를 갖는 최소한의 설정값만을 사용하면서도, 기존 알고리즘에 비해 더 높은 분석효율을 갖는 SigAlign 알고리즘을 개발했다. 이 알고리즘은 기존 데이터 처리 방법들에 비해 50배 이상의 속도로 더 정확한 결과를 제공할 수 있다. SigAlign은 질병 진단, 신약 개발, 미생물 분석 등 다양한 분야에서 활용될 수 있으며, 관련 학술분야나 산업화에도 기여할 것으로 기대된다. 연구진은 새로운 알고리즘을 활용한 솔루션을 개발할 계획이며, 누구나 자유롭게 사용할 수 있도록 전체 내용을 공개하였다.
Abstract	Professor Joohon Sung from the Graduate School of Public Health and the Interdisciplinary Program in Bioinformatics at Seoul National University, along with Ph.D. candidate Kunhyung Bahk from the same Bioinformatics program, have presented a new algorithm for genome analysis. This innovative algorithm, “SigAlign”, significantly improves upon the complexity and opacity of existing methods while conferring up to 50 times faster performance compared to the popular algorithms. This foundational technology is expected to become a core technology in various biomedical research and biotechnology applications.

	(논문 초록으로 대체) In biological sequence alignment, prevailing heuristic aligners achieve high-throughput by several approximation techniques, but at the cost of sacrificing the clarity of output criteria and creating complex parameter spaces. To surmount these challenges, we introduce 'SigAlign', a novel alignment algorithm that employs two explicit cutoffs for the results: minimum length and maximum penalty per length, alongside three affine gap penalties. Comparative analyses of SigAlign against leading database search tools (BLASTn, MMseqs2) and read mappers (BWA-MEM, bowtie2, HISAT2, minimap2) highlight its performance in read mapping and database searches. Our research demonstrates that SigAlign not only provides high sensitivity with a non-heuristic approach, but also surpasses the throughput of existing heuristic aligners, particularly for high-accuracy reads or genomes with few repetitive regions. As an open-source library, SigAlign is poised to become a foundational component to provide a transparent and customizable alignment process to new analytical algorithms, tools and pipelines in bioinformatics.
Journal Link	https://doi.org/10.1093/nar/gkae607 (Bahk, K., & Sung, J. (2024). SigAlign: an alignment algorithm guided by explicit similarity criteria. Nucleic Acids Research, gkae607.)

■ 본문

□ 학술 성과의 소개와 의의

서울대학교 연구진 (서울대학교 보건대학원 및 생물정보협동과정 교수 성주헌; 생물정보협동과정 박사과정 박건형 연구원)이 **유전체 분석을 혁신할 수 있는 새로운 원천 알고리즘을 개발**하여 공개하였다. 유전체 분석은 질병 진단과 신약 개발 등의 첨단 바이오 테크놀로지에 필수적인 핵심 기술이다. 유전체 분석은 실험적인 분석 과정뿐 아니라, 그 이후에 대규모의 생물정보 데이터 처리 단계가 필수적이다. 한 사람당 30억개의 염기서열을 분석하는 유전체 자료는PB (페타바이트) 규모에 달하여, 정확한 염기쌍을 효율적이고 빠르게 찾는 알고리즘은 매우 중요한 원천기술이다. 지금까지 이러한 원천 알고리즘은 전적으로 영-미의 기술에 의존해 왔다.

서울대 연구진이 개발한 "SigAlign" (Similarity-guided Alignment)은 현재 널리 사용되고 있는 알고리즘 (BWA-MEM, HISAT2, Bowtie2) 등에 비해 수십 배 빠른 속도로 더 정확한 결과를 도출하였으며, 표준 컴퓨팅 환경에서의 방대한 비교분석에서 초당 10만 개 이상의 데이터를 처리할 수 있는 유일한 알고리즘이었다. 또 SigAlign은 기존 알고리즘들이 수십 개의 난해한 설정값 (parameter) 옵션들을 조정해야 되고, 이런 조정과정을 거쳐도 결과를 예측하기 어려운 단점을 획기적으로 극복해, 5개의 순수한 생물학적 설정값 (불일치 점수를 제외하면 단 2개의 설정값) 만으로 유전체 분석의 모든 과정과 결과를 투명하고 예측 가능하게 만들었다는 점에서 큰 관심을 받고 있다. 본 연구는 식약처 및 연구재단 지원을 통해 수행되었다.

□ 연구개발 관련 배경 설명

인간게놈 프로젝트에서는 13년 (1990-2003)이 걸렸던 한 사람의 유전체 분석이 최근 단 하루만에 가능하게 된 것은 30억 개의 염기서열을 작은 토막 (100-400개 단위) 으로 만들어 빠르게 분석

하는 차세대염기서열 분석 (NGS, next generation sequencing) 기술과 이를 정확하고 빠르게 처리할 수 있는 생물정보학 알고리즘의 공헌이 절대적이다. 유전체 시대의 바탕이 되는 유전체 분석 기술은 실험적인 염기서열 분석 기술 (예 Illumina의 NovaSeq; Oxford Nanopore의 ElysIon 등)과 방대한 정보를 빠르고 정확하게 처리하기 위한 생물 정보 데이터 처리 알고리즘으로 나뉘어진다. 지금까지 실험적 분석기기는 물론, 자료처리 알고리즘들도 미국과 영국 등의 국가가 개발을 독점해 왔다.

기존의 생물정보 알고리즘은 방대한 자료의 처리를 위해 속도의 향상에 집중해왔다. 최근 널리 사용되는 알고리즘들은 정확도를 유지하면서 최대한 성능 (=속도)을 향상시키기 위해서, 점점 더 많은 기술적인 설정값을 추가하여 각 단계의 연산을 간소화하는 방향으로 개발되고 있다. 그러나, 현재 널리 사용되는 분석알고리즘들은 공통적으로 몇 가지의 문제점을 가지고 있다. 1) 전문가들도 설정값의 의미를 이해하기 어려워서, 새로운 분야나 특성이 다른 자료의 분석에서는 전혀 다른 결과가 나올 수 있다 (예: 불필요하게 정확도가 떨어지는 결과까지 제시하여 이후의 처리과정이 어려워짐). 2) 분석 프로그램 간의 호환성이 매우 낮아, 매우 다른 결과를 제시한다. 3) 서로 다른 결과가 왜 나왔는지, 어떤 결과가 정확한지를 알기가 굉장히 어렵다. 요약하면, 속도를 향상시키기 위한 난해성, 비호환성, 불투명성 등의 문제가 계속해서 누적되어 온 것이다. 이는 데이터 처리 과정과 결과의 신뢰성을 떨어뜨리며, 이후에 유전체 정보를 활용한 생의학적인 후속 연구개발에도 영향을 미친다.

이번에 새로 개발된 SigAlign은 가장 원천적인 유전체 분석의 단계인 염기쌍 매칭 (=alignment) 방법을 획기적으로 개선한 알고리즘으로, 모든 유전체 분석에 적용 가능한 원천기술이다. (알고리즘의 그림설명 참고) 국제 학술지에 최근 게재된 논문에서 서울대학교 연구진은 생물학적인 의미가 담긴 단순한 설정값만으로도 가장 빠르게 유전체 분석이 가능함을 입증하였고, 방대한 비교분석 실험에서도 대부분의 상황에서 더 빠르고 정확한 분석이 가능함을 검증하였다. 예를 들어, SigAlign은 장내미생물의 분석에서 다른 알고리즘에 비해 최대 50배이상 빠른 속도와 가장 높은 정확도를 보였다. 또, SigAlign은 기술적으로도 매우 효율적인 메모리 사용이 가능하도록 개발되어, 고성능 컴퓨터 (CPU 100개, 메모리 1-2 TB)가 필요했던 기존의 알고리즘과는 달리 향후 휴대폰이나 소형 IoT기기에서도 사용이 가능한 분석 알고리즘이다.

□ 연구성과의 기대 파급효과

연구진이 개발한 유전체분석 알고리즘의 정확성과 투명성 및 직관적인 사용편의성 등은 유전체 분석이 필요한 모든 분야의 분석을 최적화 시킬 수 있는 방법이 될 것으로 예상된다. 예를 들어, 인간의 질병진단, 미생물 분석, 종양분석, 단일세포 분석 등은 매우 다른 특성을 가진 영역으로, 기존의 분석방법은 최적화를 위해 매우 많은 전문인력과 시간이 필요했다.하지만 새로 개발된 알고리즘은 전문교육을 받지 않은 사람이라도 쉽게 자기 연구분야의 지식을 바탕으로 최적화가 가능하다. 따라서, 현재 유전체 기술이 적용되고 있는 분야는 물론, 향후 적용될 모든 생물학적 연구분야에 활용 가능할 것이다.

생물정보 분석은 데이터를 통해 가치를 창출하는 고부가가치 산업군으로, 생명정보기술의 “쌀”이라

고 할 수 있는 원천기술이다. 현재 해외에서는 많은 스타트업이 이 분야에 뛰어들고 있으며, 아마존과 구글 같은 빅테크 기업들도 생물 정보 분석 서비스에 진출하고 있다. 한국인의 아이디어와 기술로 만들어진 SigAlign은 이러한 글로벌 경쟁 속에서 독보적인 경쟁력을 가진 도구로, 한국의 생물 정보 산업 발전에도 큰 기여를 할 것으로 기대된다.

대규모 생물 정보 분석은 지금까지 알고리즘 보다는 주로 하드웨어의 발전을 통해 가속화되어왔다. 즉, 더 많은 CPU와 메모리를 갖는 슈퍼컴퓨터를 통해서 속도를 향상시키는 방법을 사용해 왔다. 앞에서 언급한 난해성과 비호환성, 불투명성 등의 문제 이외에도, 하드웨어 의존은 분석비용 상승과 전력소모량 증가 등의 새로운 문제를 야기시킨다. SigAlign은 새로운 개념의 알고리즘으로 속도와 컴퓨팅 효율면에서도 뛰어난 성능을 가지고 있어서, 그만큼 하드웨어 의존성을 줄일 수 있고 비용을 절감할 수 있다.

연구진은 이번에 발표한 SigAlign 알고리즘을 특허화하기 보다는 공개-개방해서 많은 사람들이 자유롭게 사용할 수 있도록 하였다. 이는 더 많은 사람들이 연구성과를 활용하여 더 좋은 연구를 효율적으로 수행하게 하기 위한 선택이다. 그러나, 알고리즘이 공개된다고 해도, 이를 활용한 다양한 솔루션의 개발은 난이도가 높은 작업으로, 연구진이 지속적으로 주도권을 가지고 새로운 응용 분야를 개척하고 발전시킬 것이다.

※ 연구 이야기

□ 연구를 시작한 계기

연구진은 수년간 유전체 연구에 기존의 분석 알고리즘들을 활용하는 도중, 프로그램마다 결과의 차이가 현저할 뿐 아니라, 그 어떤 것도 결과에 대한 명확한 “기준”을 제시하지 못한다는 사실을 알게 되었다. 인간 유전체 분석에 사용한 알고리즘을 인간의 장내 미생물에 사용하면 예상과 전혀 다른 결과를 보여주는데, 그 이유를 파악하는 것은 원저자와 상의를 해봐도 거의 불가능에 가까웠다. 이러한 문제의식이 쌓이다가, 아예 새로운 알고리즘을 개발해 보려는 생각으로 귀결되었고, 3년에 걸친 집중적인 작업 끝에 성과를 낼 수 있었다.

□ 알고리즘의 활용 가능성과 잠재적 응용 분야

SigAlign은 DNA나 RNA를 실험적으로 분석한 데이터라면 어디에든 적용 가능하다. 하지만, 이 중에서도 특히 SigAlign은 새로운 분석 분야에서 강점을 가진다. 유전체 분석은 아직 특화된 툴이 전혀 개발되지 않은 분야도 굉장히 많다. 연구진의 성과는 새로운 문제에 대한 최적화에 특화된 알고리즘이기 때문에 이러한 미개척 분야의 선두 주자가 될 수 있다. 또한, 매우 작은 하드웨어에서도 동작 가능하다는 장점으로 인해 SigAlign은 대규모 컴퓨팅 장비 없이 사용 가능한 유일한 솔루션이 될 수도 있다. 예를 들어 Oxford Nanopore Technology (ONT)는 최초로 손바닥 만한 실험분석기기인 MinION을 만들어 아프리카의 현장에서 에볼라 바이러스를 진단하는 성과를 보였는데, SigAlign은 이처럼 장비가 제한된 상황에서 최적의 솔루션이 될 수 있다.

□ 알고리즘의 공개와 실용화 전망

SigAlign 분석방법은 공개되어 누구나 사용할 수 있다. 알고리즘의 코드는 국제적인 전문가들의 공개 검증을 위해 연구실 및 연구자의 GitHub에 올라와 있으며, 기초 컴퓨터 지식이 있는 사람은 실제 분석에 자유롭게 활용 가능하다. (저장소 링크- https://github.com/ghbd-snu, https://github.com/baku4/sigalign)

연구진은 SigAlign을 실제 솔루션으로 개발할 계획도 가지고 있다. 현재의 연구개발 내용만으로도 솔루션 개발이 가능하지만, 이를 위해서는 개발된 핵심 알고리즘 이외에도 사용자를 고려한 환경(API 등의 인터페이스)을 포함한 실용 서비스 개발을 위한 인력과 비용이 추가로 필요하다. 따라서 현재는 실용화보다는 종양 분석이나 단백질 서열 분석, 정확성이 낮고 긴 유전 정보 분석(30만 개 이상의 염기서열) 분야를 위한 알고리즘의 추가 개선에 더 집중하고 있다.

SigAlign의 공개는 독점보다는 더 많은 사람들이 사용할 수 있도록 하기 위한 선택이었다. 이론적으로는 다른 사람이 이 아이디어를 가지고 상용화할 수도 있지만, 현실적으로 이는 매우 어렵다. 새로운 알고리즘의 내용을 완벽하게 이해하는 것은 수십 년간 분석 분야에 종사해온 전문가들에게도 쉽지 않은 일이다. 또한, 공개 코드라 할지라도 사용자는 저작자의 사용 규정을 반드시 준수해야 하기 때문에, 향후 이 알고리즘을 활용한 솔루션의 개발은 연구진이 주도할 것으로 예상된다.