Analisis Pensejajaran Sequence

BIOINFORMATIKA

Gadis Retno Apsari
Robiah Adawiyah
Mey Ayu Linatari
Dessy Rahmayadi
Mohammad Syaiful Pradana

BIOINFORMATIKA

ANALISIS PENSEJAJARAN SEQUENCE

Oleh:
Gadis Retno Apsari
Robiah Adawiyah
Mey Ayu Linatari
Dessy Rahmahyadi
Mohammad Syaiful Pradana

ISBN: 978-602-6715-37-1

KATA PENGANTAR

Assalamu'alaikum War. Wab.

Puji syukur kehadirat Allah SWT atas Rahman dan RohimNya akhirnya buku Ajar *Sequence Alignment* Berbasis Bioinformatika ini dapat terselesaikan, Shalawat dan salam semoga senantiasa terlimpahkan pada Nabi Muhammad SAW keluarga dan sahabatnya.

Genetika merupakan salah satu materi yang sulit dimengerti oleh sebagian besar siswa karena konsep genetika bersifat esoteric dan abstrak, yang meliputi objek-objek mikroskopik dan proses-proses diluar pengalaman siswa sehari-hari. Selain itu, model pembelajaran mata pelajaran biologi untuk sekolah menengah atas paling sering menggunakan kelas untuk teori dan laboratorium basah untuk prakteknya. Namun, pembelajaran materi genetika dapat dilakukan di laboratorium kering seperti laboratorium komputer. Pembelajaran materi genetika ini dapat dilaksanakan dilaboratorium komputer dengan dasar Bioinformatika.

Buku ini memberikan penjelasan bagaimana siswa mempelajari materi genetika dengan dasar Bioinformatika mulai dari cara pengambilan data genetik dari web gen NCBI (*National Center for Biotechnoogy Information*), penjelasan mengenai DNA, RNA dan protein, hingga aplikasi untuk pensejajaran sequence. Buku ini sesuai dibaca bagi kalangan siswa sekolah menengah atas untuk lebih dalam mempelajari genetika.

Penulis mengucapkan terima kasih kepada RistekDikti yang telah memfasilitasi dalam penyusunan buku ini melalui program kreativitas mahasiswa pengabdian kepada masyarakat (PKMM) tahun 2018. Penulis menyadari tulisan ini masih terdapat beberapa kekurangan, untuk itu diharapkan saran dan kritik yang membangun dari pembaca. Semoga tulisan ini bermanfaat.

Wassalamu'alaikum War. Wab.

Tim Penulis

DAFTAR ISI

Bab 1

Bioinformatika

1.1 Pendahuluan

Perkembangan teknologi informasi dalam berbagai bidang ilmu telah mengembangkan bidang ilmu yang bersangkutan. Berbagai kajian baru muncul sejalan dengan perkembangan teknologi informasi. Aplikasi teknologi informasi dalam bidang biologi molekuler telah melahirkan bidang ilmu Bioinformatika. Kajian Bioinformatika ini tak lepas dari perkembangan biologi molekul modern yang ditandai dengan kemampuan manusia untuk memahami genom yaitu infomasi genetic yang menentukan sifat setiap makhluk hidup yang disandi dalam bentuk pita molekul DNA (*Asam Deoksiribonukleat*). Kemampuan untuk memahami dan memanipulasi kode genetic DNA ini sangat didukung oleh teknologi informasi melalui perangkat lunak maupun keras.

Kelahiran Bioinformatika modern tak lepas dari perkembangan bioteknologi di era tahun 70-an, dimana seorang ilmuwan Amerika Serikat melakukan inovasi dalam mengembangkan teknologi DNA rekombinan. Berkat penemuan ini lahirlah perusahaan bioteknologi pertama didunia, yaitu Genentech di AS, yang kemudian memperoduksi protein hormone insulin dan bakteri yang dibutuhkan oleh penderita diabetes. Selama ini insulin hanya bias didapatkan dalam jumlah sangat terbatas dari organ pankreas sapi.

Bioteknologi modern ditandai dengan kemampuan pada manipulasi DNA. Rantai/sekuen DNA yang mengkode protein disebut gen. Gen ditranskripsikan menjadi mRNA, kemudian mRNA ditranslasikan menjadi protein. Protein sebagai produk akhir bertugas menunjang seluruh proses kehidupan, antara lain sebagai katalis reaksi biokimia dalam tubuh (disebut enzim), berperan serta dalam sistem pertahanan tubuh melawan virus, parasit dan lain-lain (disebut antibodi), menyusun struktur tubuh dari ujung kaki (otot terbentuk dari protein actin, myosin, dan sebagainya) sampai ujung rambut (rambut tersusun dari protein keratin), dan lain-lain. Arus informasi, DNA -> RNA -> Protein, inilah yang disebut sentral dogma dalam biologi molekul.

Sekuen DNA satu organisme, yaitu pada sejenis virus yang memiliki kurang lebih 5.000 nukleotida/molekul DNA atau sekitar 11 gen, berhasil dibaca secara menyeluruh pada tahun 1977. Sekuen seluruh DNA manusia terdiri dari 3 milyar nukleotida yang menyusun 100.000 gen dapat dipetakan dalam waktu 3 tahun. Saat ini terdapat milyaran data nukleotida yang tersimpan dalam database DNA, GenBank di AS yang didirikan tahun 1982. Di Indonesia, ada Lembaga Biologi Molekul Eijkman yang terletak di Jakarta. Desakan kebutuhan untuk mengumpulkan, menyimpan dan menganalisa data-data biologis dari database DNA, RNA maupun protein inilah yang semakin memacu perkembangan kajian Bioinformatika.

1.2 Pengertian Bioinformatika

Pada bagian pendahuluan kita telah diberikan gambaran sekilas mengenai perkembangan dan apa yang dapat diberikan oleh Bioinformatika. Bagian berikut ini akan membahas lebih detail tentang Bioinformatika.

Bioinformatika adalah ilmu yang mempelajari penerapan teknik komputational untuk mengelola dan menganalisis informasi biologis. Bidang ini mencangkup penerapan metod-metode matematika, statistika dan informatika untuk memecahkan masalah-masalah biologis, terutama dengan menggunakan sekuens DNA dan asam amino serta informasi yang berkaitan dengannya. Contoh topic utama bidang ini meliputi basis data untuk mengelola informasi biologis, pensejajaran sekuens (*sequence alignment*), prediksi struktur untuk meramalkan bentuk struktur protein maupun struktur seunder RNA, analisis filogenetik, dan analisis ekspresi gen.

Secara umum, Bioinformatika dapat digambarkan sebagai: segala bentuk penggunaan komputer dalam menangani informasi-informasi biologi. Dalam prakteknya, definisi yang digunakan oleh kebanyakan orang bersifat lebih terperinci. Bioinformatika menurut kebanyakan orang adalah satu sinonim dari komputasi biologi molekul (penggunaan komputer dalam menandai karakterisasi dari komponen-komponen molekul dari makhluk hidup).

1. Bioinformatika "klasik"

Sebagian besar ahli Biologi mengistilahkan mereka sedang melakukan Bioinformatika' ketika mereka sedang menggunakan komputer untuk menyimpan, melihat atau mengambil data, menganalisa atau

memprediksi komposisi atau struktur dari biomolekul. Ketika kemampuan komputer menjadi semakin tinggi maka proses yang dilakukan dalam Bioinformatika dapat ditambah dengan melakukan simulasi. Yang termasuk biomolekul diantaranya adalah materi genetik dari manusia -- asam nukleat--dan produk dari gen manusia, yaitu protein. Hal-hal diataslah yang merupakan bahasan utama dari Bioinformatika "klasik", terutama berurusan dengan analisis sekuen (*sequence analysis*).

Definisi Bioinformatika menurut Fredj Tekaia dari Institut Pasteur [TEKAIA2004] adalah: "metode matematika, statistik dan komputasi yang bertujuan untuk menyelesaikan masalah-masalah biologi dengan menggunakan sekuen DNA dan asam amino dan informasi-informasi yang terkait dengannya."

Dari sudut pandang Matematika, sebagian besar molekul biologi mempunyai sifat yang menarik, yaitu molekul-molekul tersebut adalah polymer; rantai-rantai yang tersusun rapi dari modul-modul molekul yang lebih sederhana, yang disebut monomer. Monomer dapat dianalogikan sebagai bagian dari bangunan, dimana meskipun bagianbagian tersebut berbeda warna dan bentuk, namun semua memiliki ketebalan yang sama dan cara yang sama untuk dihubungkan antara yang satu dengan yang lain.

Monomer yang dapat dikombinasi dalam satu rantai ada dalam satu kelas umum yang sama, namun tiap jenis monomer dalam kelas tersebut mempunyai karakteristik masing-masing yang terdefinisi dengan baik.

Beberapa molekul-molekul monomer dapat digabungkan bersama membentuk sebuah entitas yang berukuran lebih besar, yang disebut *macromolecule*. *Macromolecule* dapat mempunyai informasi isi tertentu yang menarik dan sifat-sifat kimia tertentu.

Berdasarkan skema di atas, monomer-monomer tertentu dalam *macromolecule* dari DNA dapat diperlakukan secara komputasi sebagai huruf-huruf dari alfabet, yang diletakkan dalam sebuah aturan yang telah diprogram sebelumnya untuk membawa pesan atau melakukan kerja di dalam sel. Proses yang diterangkan di atas terjadi pada tingkat molekul di dalam sel. Salah satu cara untuk mempelajari proses tersebut selain dengan mengamati dalam laboratorium biologi yang sangat khusus adalah dengan menggunakan Bioinformatika sesuai dengan definisi "klasik" yang telah disebutkan di atas.

2. Bioinformatika "baru"

Salah satu pencapaian besar dalam metode Bioinformatika adalah selesainya proyek pemetaan genom manusia (*Human Genome Project*). Selesainya proyek raksasa tersebut menyebabkan bentuk dan prioritas dari riset dan penerapan Bioinformatika berubah. Secara umum dapat dikatakan bahwa proyek tersebut membawa perubahan besar pada sistem hidup kita, sehingga sering disebutkan terutama oleh ahli biologi bahwa kita saat ini berada di masa pascagenom.

Selesainya proyek pemetaan genom manusia ini membawa beberapa perubahan bagi Bioinformatika, diantaranya:

Setelah memiliki beberapa genom yang utuh maka kita dapat mencari perbedaan dan persamaan di antara gen-gen dari spesies yang berbeda. Dari studi perbandingan antara gen-gen tersebut dapat ditarik kesimpulan tertentu mengenai spesies-spesies dan secara umum mengenai evolusi. Jenis cabang ilmu ini sering disebut sebagai perbandingan genom (*comparative genomics*).

Sekarang ada teknologi yang didisain untuk mengukur jumlah relatif dari kopi/cetakan sebuah pesan genetik (level dari ekspresi genetik) pada beberapa tingkatan yang berbeda pada perkembangan atau penyakit atau pada jaringan yang berbeda. Teknologi tersebut, contohnya seperti *DNA microarrays* akan semakin penting.

Akibat yang lain, secara langsung, adalah cara dalam skala besar untuk mengidentifikasi fungsi-fungsi dan keterkaitan dari gen (contohnya metode *yeast twohybrid*) akan semakin tumbuh secara signifikan dan bersamanya akan mengikuti Bioinformatika yang berkaitan langsung dengan kerja fungsi genom (*functional genomics*).

Akan ada perubahan besar dalam penekanan dari gen itu sendiri ke hasil-hasil dari gen. Yang pada akhirnya akan menuntun ke: usaha untuk mengkatalogkan semua aktivitas dan karakteristik interaksi antara semua hasil-hasil dari gen (pada manusia) yang disebut proteomics; usaha untuk mengkristalisasi dan memprediksikan struktur-struktur dari semua protein (pada manusia) yang disebut structural genomics.

Apa yang disebut orang sebagai *research informatics* atau *medical informatics*, manajemen dari semua data eksperimen biomedik yang berkaitan dengan molekul atau pasien tertentu mulai dari spektroskop

massal, hingga ke efek samping klini akan berubah dari semula hanya merupakan kepentingan bagi mereka yang bekerja di perusahaan obat-obatan dan bagian TI Rumah Sakit akan menjadi jalur utama dari biologi molekul dan biologi sel, dan berubah jalur dari komersial dan klinikal ke arah akademis.

Dari uraian di atas terlihat bahwa Bioinformatika sangat mempengaruhi kehidupan manusia, terutama untuk mencapai kehidupan yang lebih baik. Penggunaan komputer yang notabene merupakan salah satu keahlian utama dari orang yang bergerak dalam TI merupakan salah satu unsur utama dalam Bioinformatika, baik dalam Bioinformatika "klasik" maupun Bioinformatika "baru".

1. Biophysics

Biologi molekul sendiri merupakan pengembangan yang lahir dari *biophysics*. *Biophysics* adalah sebuah bidang interdisipliner yang mengaplikasikan teknik-teknik dari ilmu Fisika untuk memahami struktur dan fungsi biologi (*British Biophysical Society*).

2. Computational Biology

Computational biology merupakan bagian dari Bioinformatika (dalam arti yang paling luas) yang paling dekat dengan bidang Biologi umum klasik. Fokus dari computational biology adalah gerak evolusi, populasi, dan biologi teoritis daripada biomedis dalam molekul dan sel. Tak dapat dielakkan bahwa Biologi Molekul cukup penting dalam *computational biology*, namun itu bukanlah inti dari disiplin ilmu ini. Pada penerapan *computational biology*, model-model statistika untuk fenomena biologi lebih disukai dipakai dibandingkan dengan model sebenarnya. Dalam beberapa hal cara tersebut cukup baik mengingat pada kasus tertentu eksperimen langsung pada fenomena biologi cukup sulit.

Tidak semua dari *computational biology* merupakan Bioinformatika, seperti contohnya Model Matematika bukan merupakan Bioinformatika, bahkan meskipun dikaitkan dengan masalah biologi.

3. Medical Informatics

Menurut Aamir Zakaria [ZAKARIA2004] Pengertian dari *medical informatics* adalah "sebuah disiplin ilmu yang baru yang didefinisikan sebagai pembelajaran, penemuan, dan implementasi dari struktur dan algoritma untuk meningkatkan komunikasi, pengertian dan manajemen informasi medis." Medical informatics lebih memperhatikan struktur dan algoritma untuk pengolahan data medis, dibandingkan dengan data itu sendiri. Disiplin ilmu ini, untuk alasan praktis, kemungkinan besar berkaitan dengan data-data yang didapatkan pada level biologi yang lebih "rumit" yaitu informasi dari sistem-sistem superselular, tepat pada level populasi di mana sebagian besar dari Bioinformatika lebih memperhatikan informasi dari sistem dan struktur biomolekul dan selular.

4. Chemiformatics

Cheminformatics adalah kombinasi dari sintesis kimia, penyaringan biologis, dan pendekatan *data mining* yang digunakan untuk penemuan dan pengembangan obat (*Cambridge Healthech Institute's Sixth Annual Cheminformatics conference*). Pengertian disiplin ilmu yang disebutkan di atas lebih merupakan identifikasi dari salah satu aktivitas yang paling populer dibandingkan dengan berbagai bidang studi yang mungkin ada di bawah bidang ini.

Salah satu contoh penemuan obat yang paling sukses sepanjang sejarah adalah penisilin, dapat menggambarkan cara untuk menemukan dan mengembangkan obat-obatan hingga sekarang meskipun terlihat aneh. Cara untuk menemukan dan mengembangkan obat adalah hasil dari kesempatan, observasi, dan banyak proses kimia yang intensif dan lambat. Sampai beberapa waktu yang lalu, disain obat dianggap harus selalu menggunakan kerja yang intensif, proses uji dan gagal (*trial error process*).

Kemungkinan penggunaan TI untuk merencanakan secara cerdas dan dengan mengotomatiskan proses-proses yang terkait dengan sintesis kimiawi dari komponen-komponen pengobatan merupakan suatu prospek yang sangat menarik bagi ahli kimia dan ahli biokimia. Penghargaan untuk menghasilkan obat yang dapat dipasarkan secara

lebih cepat sangatlah besar, sehingga target inilah yang merupakan inti dari *cheminformatics*.

Ruang lingkup akademis dari *cheminformatics* ini sangat luas. Contoh bidang minatnya antara lain: *Synthesis Planning, Reaction and Structure Retrieval, 3-D Structure Retrieval, Modelling, Computational Chemistry, Visualisation Tools and Utilities*.

5. Genomics

Genomics adalah bidang ilmu yang ada sebelum selesainya sekuen genom, kecuali dalam bentuk yang paling kasar. *Genomics* adalah setiap usaha untuk menganalisa atau membandingkan seluruh komplemen genetik dari satu spesies atau lebih. Secara logis tentu saja mungkin untuk membandingkan genom-genom dengan membandingkan kurang lebih suatu himpunan bagian dari gen di dalam genom yang representatif.

6. Mathematical Biology

Mathematical biology lebih mudah dibedakan dengan Bioinformatika daripada *computational biology* dengan Bioinformatika. *Mathematical biology* juga menangani masalah-masalah biologi, namun metode yang digunakan untuk menangani masalah tersebut tidak perlu secara numerik dan tidak perlu diimplementasikan dalam *software* maupun *hardware*. Bahkan metode yang dipakai tidak perlu "menyelesaikan" masalah apapun; dalam *mathematical biology* bisa dianggap beralasan untuk mempublikasikan sebuah hasil yang hanya menyatakan bahwa suatu masalah biologi berada pada kelas umum tertentu.

Menurut Alex Kasman [KASMAN2004] Secara umum *mathematical biology* melingkupi semua ketertarikan teoritis yang tidak perlu merupakan sesuatu yang beralgoritma, dan tidak perlu dalam bentuk molekul, dan tidak perlu berguna dalam menganalisis data yang terkumpul.

7. Proteomics

Istilah *proteomics* pertama kali digunakan untuk menggambarkan himpunan dari protein-protein yang tersusun (*encoded*) oleh genom. Ilmu yang mempelajari *proteome*, yang disebut *proteomics*, pada saat ini tidak hanya memperhatikan semua protein di dalam sel yang diberikan, tetapi juga himpunan dari semua bentuk isoform dan modifikasi dari semua protein, interaksi diantaranya, deskripsi struktural dari proteinprotein dan kompleks-kompleks orde tingkat tinggi dari protein, dan mengenai masalah tersebut hampir semua pasca genom.

Michael J. Dunn [DUNN2004], Pemimpin Redaksi dari *Proteomics* mendefiniskan kata "*proteome*" sebagai: "*The PROTEin complement of the genOME*". Dan mendefinisikan *proteomics* berkaitan dengan: "studi kuantitatif dan kualitatif dari ekspresi gen di level dari protein-protein fungsional itu sendiri". Yaitu: "sebuah antarmuka antara biokimia protein dengan biologi molekul".

Mengkarakterisasi sebanyak puluhan ribu protein-protein yang dinyatakan dalam sebuah tipe sel yang diberikan pada waktu tertentu apakah untuk mengukur berat molekul atau nilai-nilai isoelektrik protein-protein tersebut melibatkan tempat penyimpanan dan perbandingan dari data yang memiliki jumlah yang sangat besar, tak terhindarkan lagi akan memerlukan Bioinformatika.

8. Pharmacogenetics

Pharmacogenomics adalah aplikasi dari pendekatan genomik dan teknologi pada identifikasi dari target-target obat. Contohnya meliputi menjaring semua genom untuk penerima yang potensial dengan menggunakan cara Bioinformatika, atau dengan menyelidiki bentuk pola dari ekspresi gen di dalam baik patogen maupun induk selama terjadinya infeksi, atau maupun dengan memeriksa karakteristik pola-pola ekspresi yang ditemukan dalam tumor atau contoh dari pasien untuk kepentingan diagnose (kemungkinan untuk mengejar target potensial terapi kanker).

Istilah *pharmacogenomics* digunakan lebih untuk urusan yang lebih "trivial" tetapi dapat diargumentasikan lebih berguna dari

aplikasi pendekatan Bioinformatika pada pengkatalogan dan pemrosesan informasi yang berkaitan dengan ilmu Farmasi dan Genetika, untuk contohnya adalah pengumpulan informasi pasien dalam database.

1.3 Teknologi dan Penerapan Bioinformatika

1. Program-program Bioinformatika

Sehari-harinya bionformatika dikerjakan dengan menggunakan program pencari sekuen (*sequence search*) seperti BLAST, program analisa sekuen (*sequence analysis*) seperti EMBOSS dan paket Staden, program prediksi struktur seperti THREADER atau PHD atau program *imaging/modelling* seperti RasMol dan WHATIF. Contoh-contoh di atas memperlihatkan bahwa telah banyak program pendukung yang mudah di akses dan dipelajari untuk menggunakan Bioinformatika

2. Teknologi Bioinformatika Secara Umum

Pada saat ini banyak pekerjaan Bioinformatika berkaitan dengan teknologi *database*. Penggunaan database ini meliputi baik tempat penyimpanan database "umum" seperti GenBank atau PDB maupun database "pribadi", seperti yang digunakan oleh grup riset yang terlibat dalam proyek pemetaan gen atau database yang dimiliki oleh perusahaan-perusahaan bioteknologi. Konsumen dari data Bioinformatika menggunakan platform jenis komputer dalam kisaran: mulai dari mesin UNIX yang lebih canggih dan kuat yang dimiliki oleh pengembang dan kolektor hingga ke mesin Mac yang lebih bersahabat yang sering ditemukan menempati laboratorium ahli biologi yang tidak suka komputer.

Database dari sekuen data yang ada dapat digunakan untuk mengidentifikasi homolog pada molekul baru yang telah dikuatkan dan disekuenkan di laboratorium. Dari satu nenek moyang mempunyai sifat-sifat yang sama, atau homology, dapat menjadi indikator yang sangat kuat di dalam Bioinformatika.

Setelah informasi dari database diperoleh, langkah berikutnya adalah menganalisa data. Pencarian database umumnya berdasarkan pada hasil

alignment / pensejajaran sekuen, baik sekuen DNA maupun protein. Kegunaan dari pencarian ini adalah ketika mendapatkan suatu sekuen DNA/protein yang belum diketahui fungsinya maka dengan membandingkannya dengan yang ada dalam database bisa diperkirakan fungsi daripadanya. Salah satu perangkat lunak pencari database yang paling berhasil dan bias dikatakan menjadi standar sekarang adalah BLAST (Basic Local Alignment Search Tool) yang merupakan program pencarian kesamaan yang didisain untuk mengeksplorasi semua database sekuen yang diminta, baik itu berupa DNA atau protein. Program BLAST juga dapat digunakan untuk mendeteksi hubungan di antara sekuen yang hanya berbagi daerah tertentu yang memiliki kesamaan. Di bawah ini diberikan contoh beberapa alamat situs yang berguna untuk bidang biologi molekul dan genetika:

Deskripsi	Alamat
National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov/
GenBank (NIH Genetic Sequence Database)	http://www.ncbi.nlm.nih.gov/Web/Genbank/index/html
European Molecular Biology Laboratory Nucleotide Sequence	http://www.ebi.ac.uk/ebi_docs/embl_db.html
Protein Information Resource	http://www.nbrf.georgetown.edu/pir
Protein Data Bank	http://www.pdb.bnl.gov/
Restriction Enzyme Database	http://www.neb.com/rebase/rebase.html
National Center for Genome Research (NCGR)	http://www.ncgr.org/gpi/
GeneMark	http://www.dixie.biology.gatech.edu/GeneMark/eukhmm.cgi
Biotechnology Industry Organization (BIO)	http://www.bio.org

Data yang memerlukan analisa Bioinformatika dan mendapat banyak perhatian saat ini adalah data hasil DNA chip. Dengan perangkat ini dapat diketahui kuantitas dan kualitas transkripsi satu gen sehingga bias

menunjukkan gen-gen apa saja yang aktif terhadap perlakuan tertentu, misalnya timbulnya kanker, dan lain-lain.

1.4 Kondisi dan Penerapan Bioinformatika di Indonesia

1. Kondisi Bioinformatika di Indonesia

Di Indonesia, Bioinformatika masih belum dikenal oleh masyarakat luas. Hal ini dapat dimaklumi karena penggunaan komputer sebagai alat bantu belum merupakan budaya. Bahkan di kalangan peneliti sendiri, barangkali hanya para peneliti biologi molekul yang sedikit banyak mengikuti perkembangannya karena keharusan menggunakan perangkat-perangkat Bioinformatika untuk analisa data. Sementara di kalangan TI masih kurang mendapat perhatian.

Ketersediaan database dasar (DNA, protein) yang bersifat terbuka/gratis merupakan peluang besar untuk menggali informasi berharga daripadanya. Database genom manusia sudah disepakati akan bersifat terbuka untuk seluruh kalangan, sehingga dapat digali/diketahui kandidat-kandidat gen yang memiliki potensi kedokteran/farmasi. Dari sinilah Indonesia dapat ikut berperan mengembangkan Bioinformatika. Kerjasama antara peneliti bioteknologi yang memahami makna biologis data tersebut dengan praktisi TI seperti programmer, dan sebagainya akan sangat berperan dalam kemajuan Bioinformatika Indonesia nantinya.

2. Penerapan Bioinformatika di Indonesia

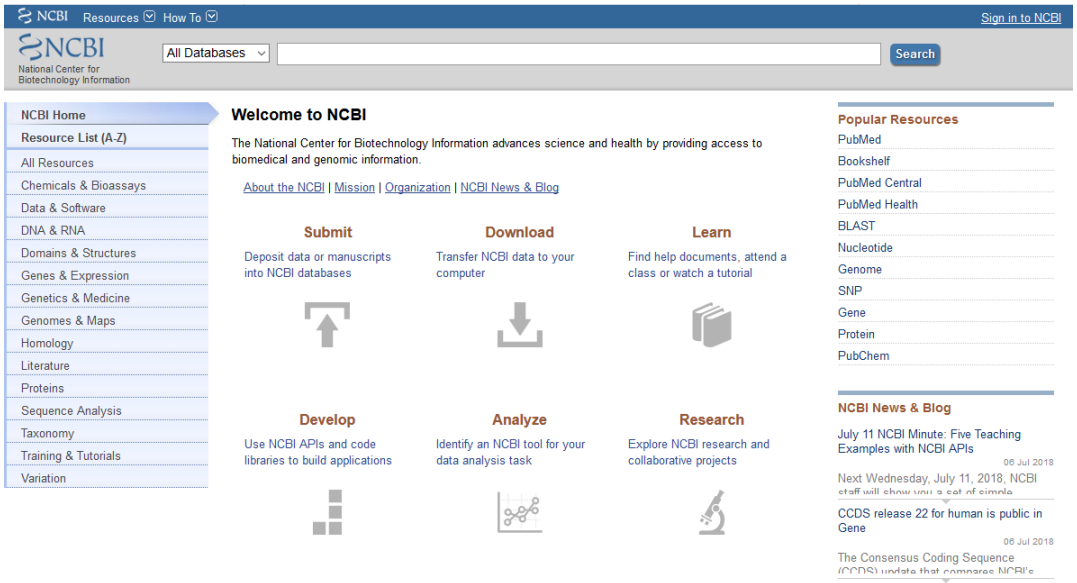
Sebagai kajian yang masih baru, Indonesia seharusnya berperan aktif dalam mengembangkan Bioinformatika ini. Paling tidak, sebagai tempat tinggal lebih dari 300 suku bangsa yang berbeda akan menjadi sumber genom, karena besarnya variasi genetiknya. Belum lagi variasi species flora maupun fauna yang berlimpah. Memang ada sejumlah pakar yang telah mengikuti perkembangan Bioinformatika ini, misalnya para peneliti dalam Lembaga Biologi Molekul Eijkman. Mereka cukup berperan aktif dalam memanfaatkan kajian Bioinformatika. Bahkan, lembaga ini telah memberikan beberapa sumbangan cukup berarti, antara lain: Deteksi kelainan janin, penembangan vaksin hepatitis B rekombinan, meringankan kelumpuhan dengan rekayasa RNA.

Bab 2

NCBI (*National Center for Biotechnology Information*)

2.1 Pengenalan NCBI *(National Center for Biotechnology Information)*

NCBI merupakan server yang memuat data base tentang informasi kesehatan dan bioteknologi. Data base terus menerus di update sesuai dengan penemuan-penemuan terkini yang menyangkut DNA, Protein, Senyawa aktif dan taksonomi. Disamping data base, NCBI juga menyediakan berbagai macam software untuk analisis DNA, protein 3D, pencarian primer, pencarian *conserve* domain dan lain sebagainya. NCBI merupakan salah satu bank data gen, protein dan literature khususnya dibidang kesehatan yang terlengkap dan di acu oleh para peneliti didunia. Situs NCBI dapat diakses pada : www.ncbi.nlm.nih.gov .



Gambar 2.1 Halaman utama NCBI

2.2 DNA dan Protein

Deoxyribonucleic acid (DNA) adalah polimer asam nukleat yang tersusun secara sistematis dan merupakan pembawa informasi genetik yang diturunkan kepada keturunannya. Informasi genetik disusun dalam bentuk kodon yang berupa tiga pasang basa nukelotida

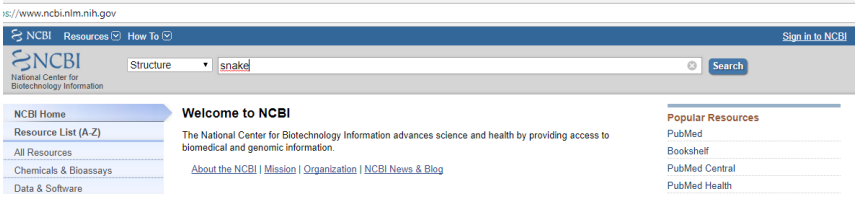
Secara struktural, DNA merupakan polimer nukleotida, di mana setiap nukelotida tersusun atas gula deoksiribosa, fosfat, dan basa. Polimer tersebut membentuk struktur dua untai heliks ganda yang disatukan oleh ikatan hydrogen antara basa-basa yang ada. Terdapat empat basa dalam DNA, yaitu adenin (A), sitosin (C), guanin (G), dan timin (T). Adenin akan membentuk dua ikatan hydrogen dengan timin, sedangkan guanin akan membentuk tiga ikatan hidrogen dengan sitosin. Kombinasi jumlah dan susunan yang terbentuk antara ikatan-ikatan basa ini memungkinkan setiap indvidu memiliki cetak biru genetik yang spesifik dibandingkan organisme lain

Protein adalah molekul penyusun tubuh kita yang terbesar setelah air. Hal ini mengindikasikan pentingnya protein dalam menopang seluruh proses kehidupan dalam tubuh. Dalam kenyataannya, memang kode genetik yang tesimpan dalam rantaian DNA digunakan untuk membuat protein, kapan, dimana dan seberapa banyak. Protein berfungsi sebagai penyimpan dan pengantar seperti hemoglobin yang memberikan warna merah pada sel darah merah kita, bertugas mengikat oksigen dan membawanya ke bagian tubuh yang memerlukan. Selain itu juga menjadi penyusun tubuh, "dari ujung rambut sampai ujung kaki", misalnya keratin di rambut yang banyak mengandung asam amino Cysteine sehingga menyebabkan bau yang khas bila rambut terbakar karena banyaknya kandungan atom sulfur di dalamnya, sampai kepada protein-protein penyusun otot kita seperti actin, myosin, titin, dsb. Kita dapat membaca teks ini juga antara lain berkat protein yang bernama rhodopsin, yaitu protein di dalam sel retina mata kita yang merubah photon cahaya menjadi sinyal kimia untuk diteruskan ke otak. Masih banyak lagi fungsi protein seperti hormon, antibodi dalam sistem kekebalan tubuh, dll.

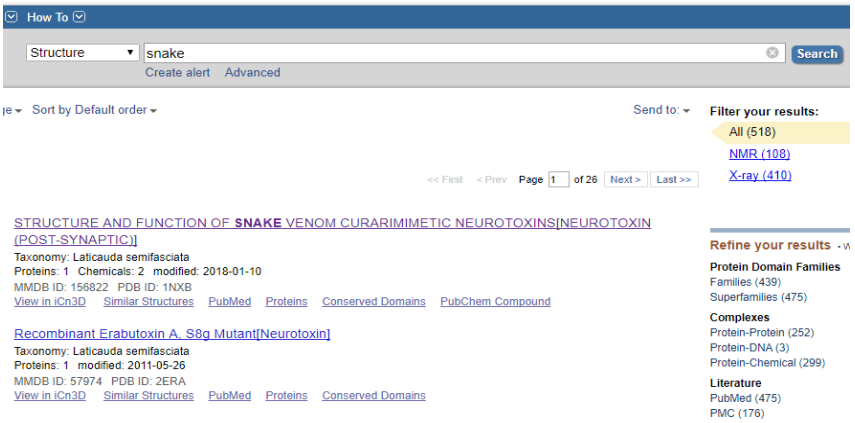
Cara mengetahui gambar protein tertentu misalnya Struktur protein Ular (*snake*). Langkah – langkahnya sebagai berikut:

1) Klik https://www.ncbi.nlm.nih.gov/
2) Pilih Structure pada menu *All Databases*,

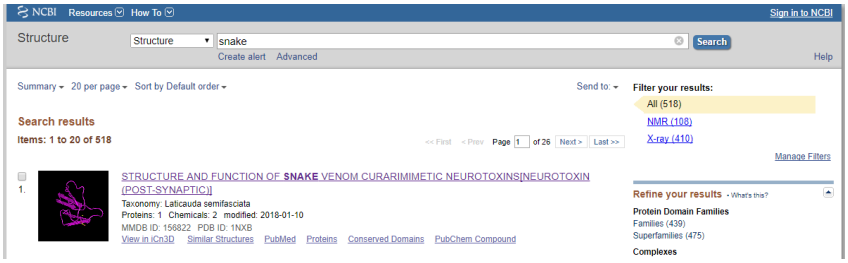
3) Ketik *snake* pada kolom kosong di samping *structure*. Klik *search.*



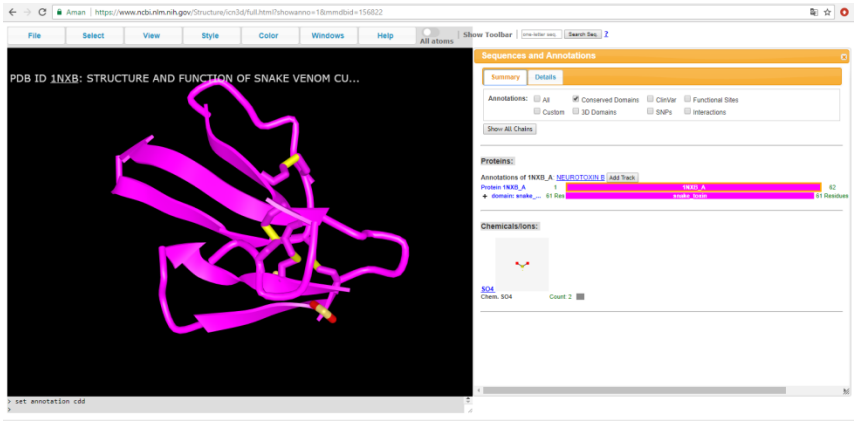
4) Pilih salah satu data dari hasil pencarian, disini kami menggunakan nomor 1



5) Klik View in iCn3D

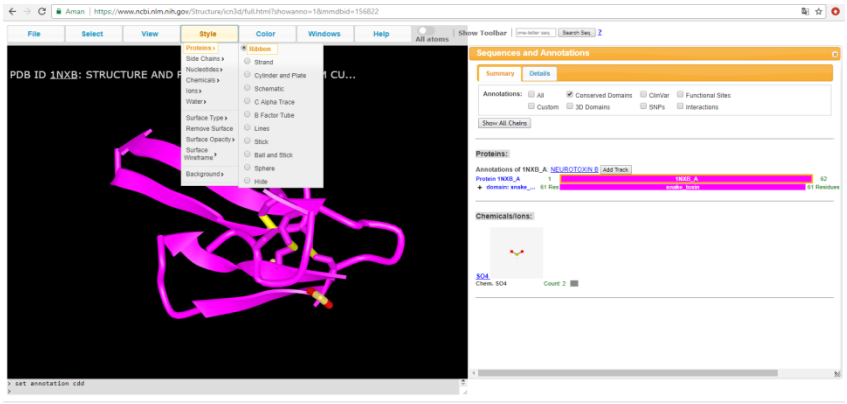


Maka akan muncul sebagai berikut

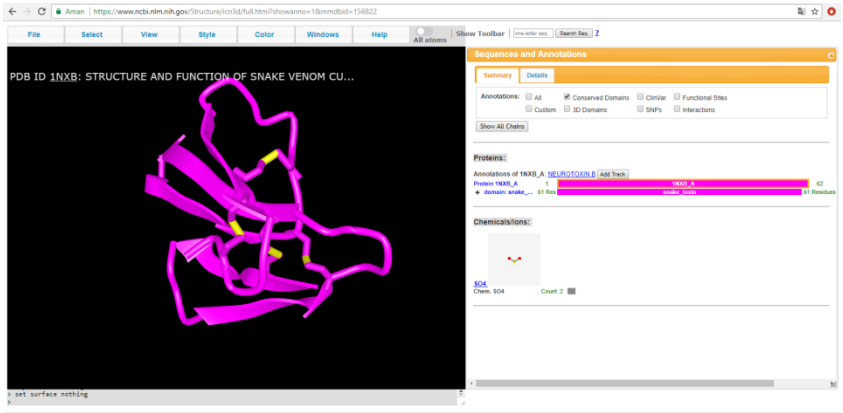


6) Klik *Style* > *Proteins* > *Ribbon* dan klik *Remove surface*

7) Untuk *side chains, nucleotides, chemicals, ions, water,* pilih opsi *Hide*



8) Maka muncul *output* seperti berikut]



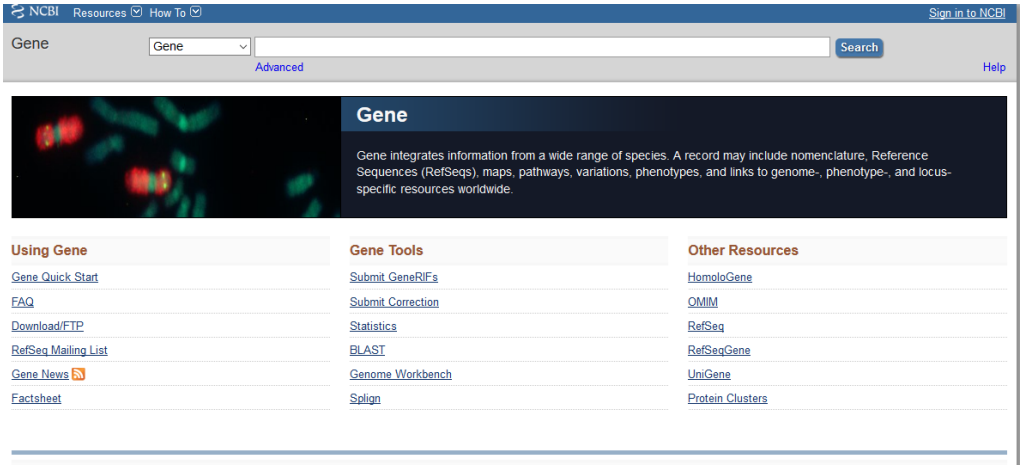
2.3 Database and Software

NCBI memiliki database dan software *(analysis tools)* yang sering digunakan untuk analisis sebagai berikut:

1. Entrez

Entrez merupakan sistem pencarian informasi dalam NCBI yang menyediakan akses terintegrasi untuk melakukan sekuensing, pemetaan (*mapping*), taksonomi dan data struktural. Entrez juga menyediakan gambaran grafis untuk mapping sekuen dan kromosom. Ciri khas dan

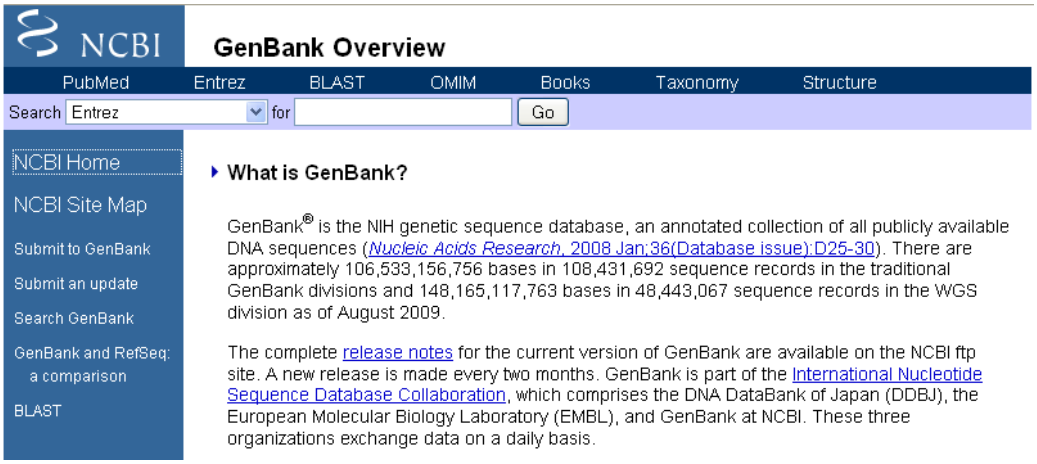
keunggulan Entrez adalah kemampuan untuk pencarian informasi terkait sekuen, struktur dan referensi. Literatur jurnal yang tersedia dapat diakses melalui PubMed. PubMed merupakan alat penghubung pencarian di web yang menyediakan akses ke lebih dari 11 juta sitasi jurnal di MEDLINE. Entrez Gene adapat diakses pada www.ncbi.nlm.nih.gov/gene .



Gambar 2.2 Entrez Gene

2. Nucleotide Database

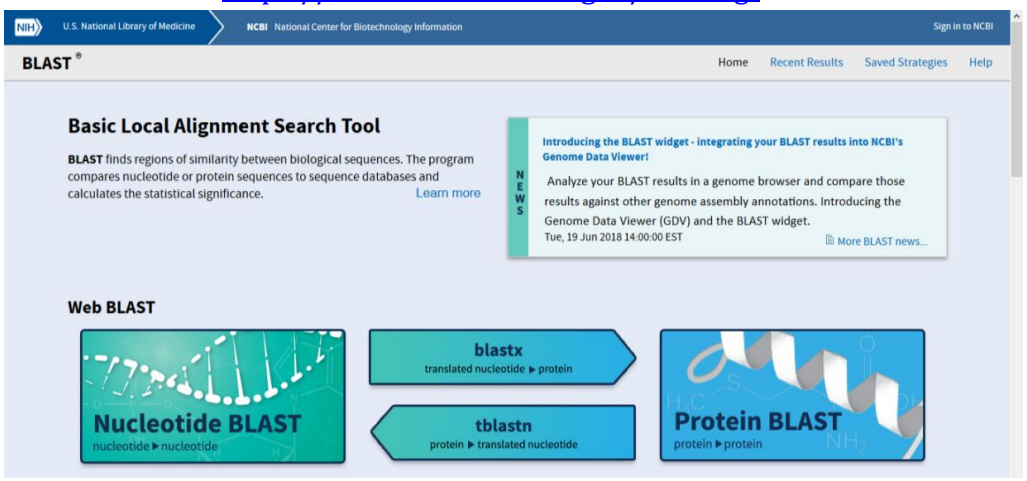
Database nukleotida merupakan suatu koleksi sekuen dari beberapa sumber, termasuk diantaranya GenBank, Reference Sequence (RefSeq), Third Party Annotation (TPA) dan Protein Data Bank (PDB). GenBank merupakan database sekuen genetik dari NIH (*National Institutes of Health*), berupa koleksi sekuen DNA yang dapat diketahui oleh publik. Database GenBank dibiayai dan didistribusikan oleh NCBI. Data sekuen dikirim ke GenBank oleh peneliti dari seluruh dunia.



Gambar 2.3 GenBank Overview

3. BLAST

BLAST (*Basic Local Alignment Search Tool*) merupakan suatu program untuk pencarian kemiripan sekuen (*sequence similarity*) dan merupakan alat dalam identifikasi gen dan karakter genetik. Blast dapat melakukan pencarian sekuen melalui perbandingan dengan database DNA dalam waktu singkat (kurang dari 15 detik). Keterangan lengkap mengenai program BLAST dapat dilihat pada : http://www.ncbi.nlm.nih.gov/blast/producttable.shtml dan BLAST dapat diakses melalui https://blast.ncbi.nlm.nih.gov/Blast.cgi.



Gambar 2.4 Halaman utama BLAST

Ada 5 program utama dalam BLAST, yaitu :

a) **nucleotide blast (blastn)** : membandingkan suatu sekuen nukleotida meragukan (*query sequence*) yang kita miliki dengan database sekuen nukleotida.

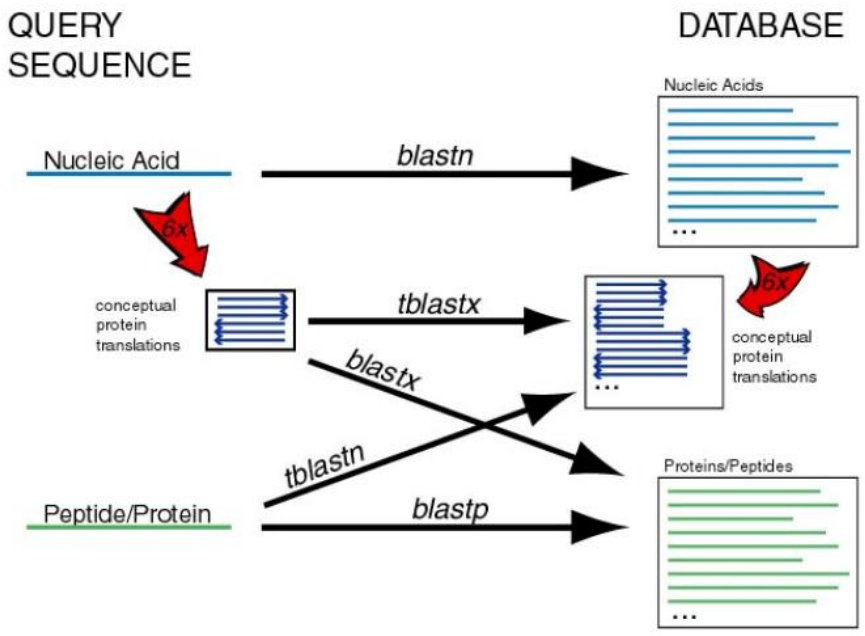
b) **protein blast (blastp)** : membandingkan suatu sekuen asam amino yang kita miliki dengan database sekuen protein.

c) **blastx** : membandingkan produk translasi konsep 6-frame sebuah sekuen nukleotida (translated nucleotide) yang kita miliki dengan database sekuen protein.

d) **tblastn** : membandingkan suatu sekuen protein yang kita miliki dengan database sekuen nukleotida yang secara dinamis ditranslasi pada semua pembacaan 6 frame.

e) **tblastx** : membandingkan suatu translasi 6 frame dari nukleotida.

Gambaran mengenai program BLAST dapat dilihat pada gambar berikut ini :



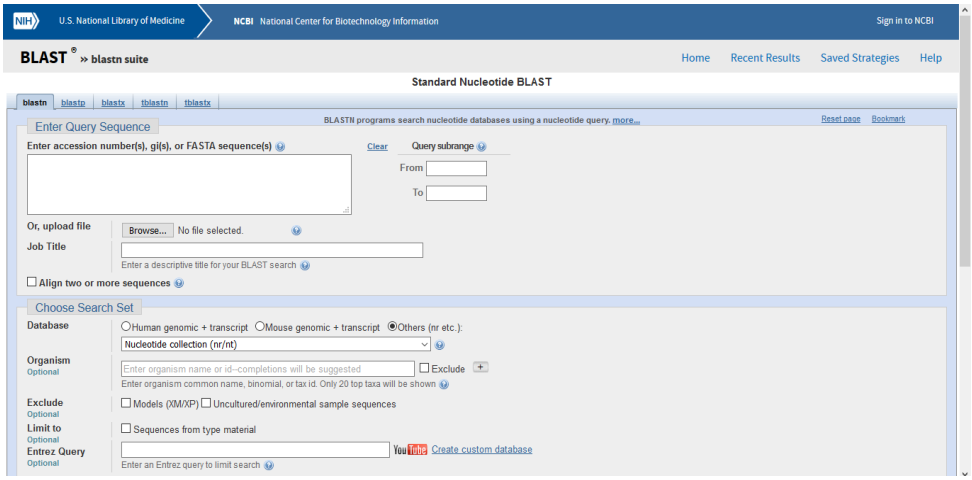
Gambar 2.5 Gambaran program utama BLAST

Disini akan diberikan contoh untuk menggunakan program utama BLAST yaitu **nucleotide blast (blastn)** dan **blastp.**

Langkah-langkah analisis menggunakan **nucleotide blast (blastn)**

Blastn dapat digunakan untuk mengidentifikasi suatu sekuen nukleotida meragukan (*query sequence*) yang kita miliki dengan database nukleotida, sehingga output yang didapat berupa identitas nukleotida tersebut, antara lain nama gen dan spesies penghasil dari sekuen lengkapnya.

1. Buka situs NCBI www.ncbi.nlm.nih.gov.
2. Pilih tool ”BLAST”, akan muncul tampilan pilihan program BLAST. Untuk mencari gen suatu sekuen nukleotida dari database nukleotida pilih ”nucleotide blast” (blastn).

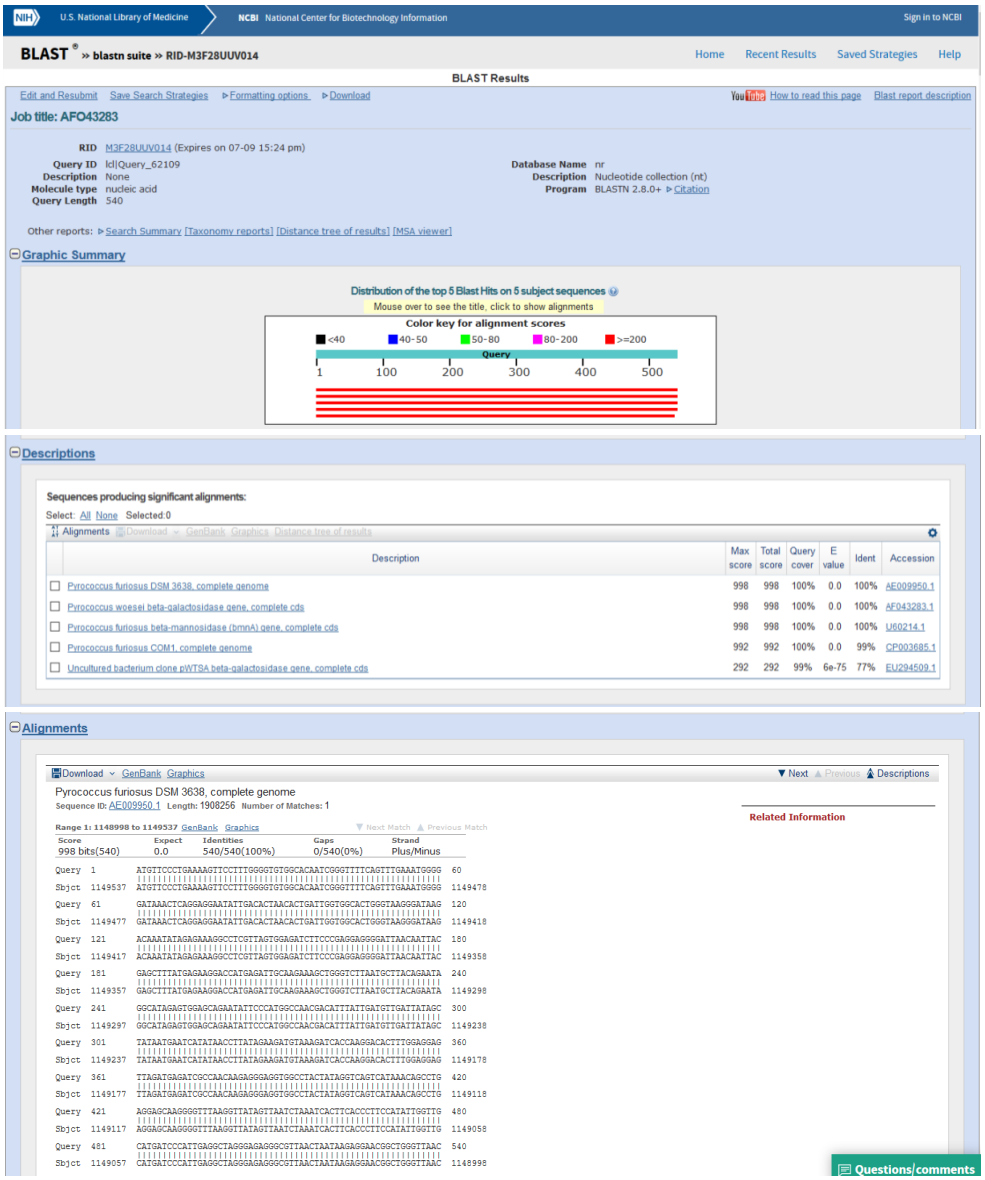


3. Setelah tampilan muncul, entri sekuen nukleotida (query) yang akan dicari; pilih setting pencarian dari database ”others” (jika belum diketahui spesiesnya); pilih program ”megablast”; klik ”BLAST” untuk memulai proses searching.
   Pada latihan/contoh digunakan sekuen nukleotida DNA berikut ini :

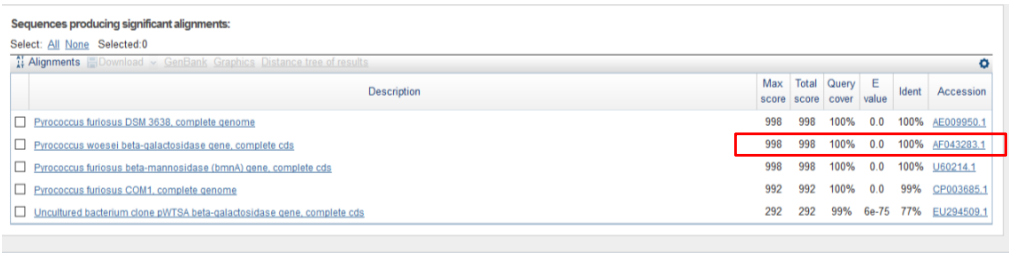
```
ATGTTCCCTGAAAAGTTCCTTTGGGGTGTGGCACAATCGGGTTTTCAGTT
TGAAATGGGGGATAAACTCAGGAGGAATATTGACACTAACACTGATTGGT
GGCACTGGGTAAGGGATAAGACAAATATAGAGAAAGGCCTCGTTAGTGGA
GATCTTCCCGAGGAGGGGATTAACAATTACGAGCTTTATGAGAAGGACCA
TGAGATTGCAAGAAAGCTGGGTCTTAATGCTTACAGAATAGGCATAGAGT
GGAGCAGAATATTCCCATGGCCAACGACATTTATTGATGTTGATTATAGC
TATAATGAATCATATAACCTTATAGAAGATGTAAAGATCACCAAGGACAC
TTTGGAGGAGTTAGATGAGATCGCCAACAAGAGGGAGGTGGCCTACTATA
```

GGTCAGTCATAAACAGCCTGAGGAGCAAGGGGTTTAAGGTTATAGTTAAT
CTAAATCACTTCACCCTTCCATATTGGTTGCATGATCCCATTGAGGCTAG
GGAGAGGGCGTTAACTAATAAGAGGAACGGCTGGGTTAAC

4. Hasil *searching* / pencarian akan didapat tampilan seperti berikut :



5. Hasil blast umumnya akan menghasilkan lebih dari satu sekuen yang bersesuaian. Pilih hasil dengan skor paling tinggi dan query coverage mendekati 100%.



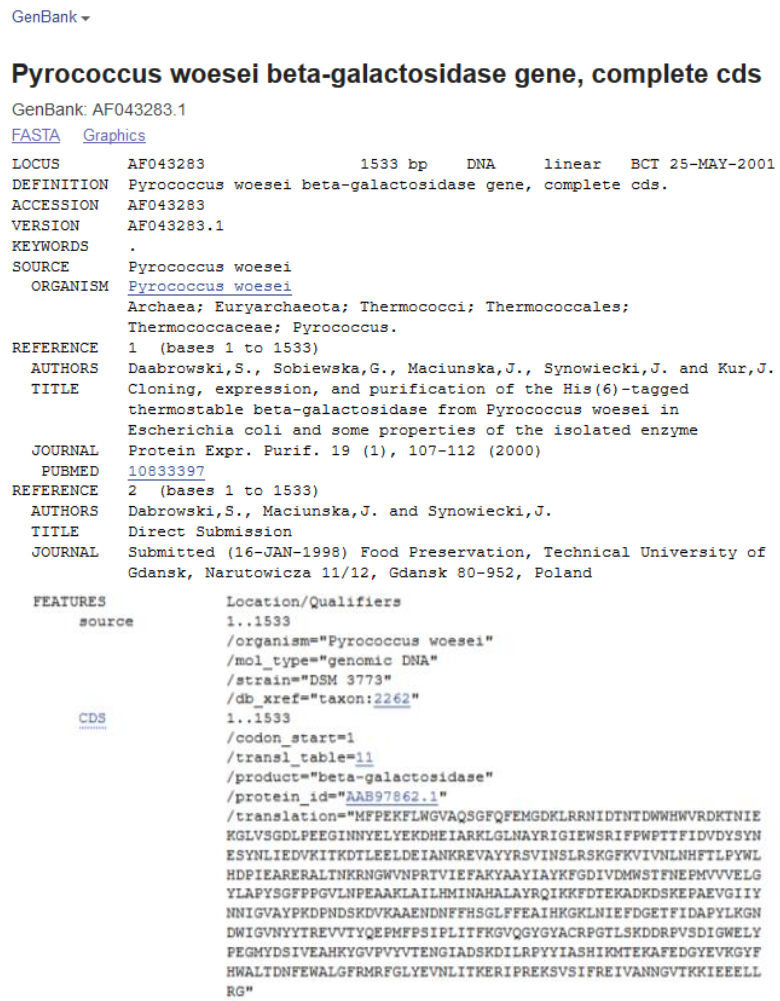
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Pyrococcus furiosus DSM 3638, complete genome	998	998	100%	0.0	100%	AE009950.1
Pyrococcus woesei beta-galactosidase gene, complete cds	998	998	100%	0.0	100%	AF043283.1
Pyrococcus furiosus beta-mannosidase (bmnA) gene, complete cds	998	998	100%	0.0	100%	U60214.1
Pyrococcus furiosus COM1, complete genome	992	992	100%	0.0	99%	CP003685.1
Uncultured bacterium clone pWTSA beta-galactosidase gene, complete cds	292	292	99%	6e-75	77%	EU294509.1

6. Klik “Accession” gen terpilih (hasil blastn) untuk keterangan lebih lanjut, (nucleotide origin dan CDS-nya).



GenBank

Pyrococcus woesei beta-galactosidase gene, complete cds

GenBank: AF043283.1

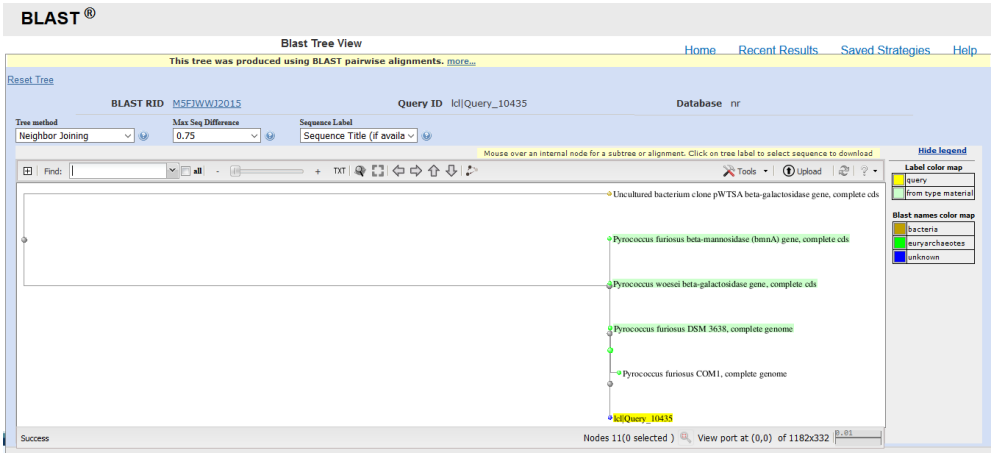
FASTA Graphics

```
LOCUS       AF043283               1533 bp    DNA     linear   BCT 25-MAY-2001
DEFINITION  Pyrococcus woesei beta-galactosidase gene, complete cds.
ACCESSION   AF043283
VERSION     AF043283.1
KEYWORDS    .
SOURCE      Pyrococcus woesei
  ORGANISM  Pyrococcus woesei
            Archaea; Euryarchaeota; Thermococci; Thermococcales;
            Thermococcaceae; Pyrococcus.
REFERENCE   1  (bases 1 to 1533)
  AUTHORS   Daabrowski,S., Sobiewska,G., Maciunska,J., Synowiecki,J. and Kur,J.
  TITLE     Cloning, expression, and purification of the His(6)-tagged
            thermostable beta-galactosidase from Pyrococcus woesei in
            Escherichia coli and some properties of the isolated enzyme
  JOURNAL   Protein Expr. Purif. 19 (1), 107-112 (2000)
   PUBMED   10833397
REFERENCE   2  (bases 1 to 1533)
  AUTHORS   Dabrowski,S., Maciunska,J. and Synowiecki,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-JAN-1998) Food Preservation, Technical University of
            Gdansk, Narutowicza 11/12, Gdansk 80-952, Poland

  FEATURES             Location/Qualifiers
     source            1..1533
                       /organism="Pyrococcus woesei"
                       /mol_type="genomic DNA"
                       /strain="DSM 3773"
                       /db_xref="taxon:2262"
     CDS               1..1533
                       /codon_start=1
                       /transl_table=11
                       /product="beta-galactosidase"
                       /protein_id="AAB97862.1"
                       /translation="MFPEKFLWGVAQSGFQFEMGDKLRRNIDTNTDWWHWVRDKTNIE
                       KGLVSGDLPEEGINNYELYEKDHEIARKLGLNAYRIGIEWSRIFPWPTTFIDVDYSYN
                       ESYNLIEDVKITKDTLEELDEIANKREVAYYRSVINSLRSKGFKVIVNLNHFTLPYWL
                       HDPIEARERALTNKRNGWVNPRTVIEFAKYAAYIAYKFGDIVDMWSTFNEPMVVVELG
                       YLAPYSGFPPGVLNPEAAKLAILHMINAHALAYRQIKKFDTEKADKDSKEPAEVGIIY
                       NNIGVAYPKDPNDSKDVKAAENDNFFHSGLFFEAIHKGKLNIEFDGETFIDAPYLKGN
                       DWIGVNYYTREVVTYQEPMFPSIPLITFKGVQGYGYACRPGTLSKDDRPVSDIGWELY
                       PEGMYDSIVEAHKYGVPVYVTENGIADSKDILRPYYIASHIKMTEKAFEDGYEVKGYF
                       HWALTDNFEWALGFRMRFGLYEVNLITKERIPREKSVSIFREIVANNGVTKKIEEELL
                       RG"
```

```
ORIGIN
        1 atgttccctg aaaagttcct ttggggtgtg gcacaatcgg gttttcagtt tgaaatgggg
       61 gataaactca ggaggaatat tgacactaac actgattggt ggcactgggt aagggataag
      121 acaaatatag agaaaggcct cgttagtgga gatcttcccg aggaggggat taacaattac
      181 gagctttatg agaaggacca tgagattgca agaaagctgg gtcttaatgc ttacagaata
      241 ggcatagagt ggagcagaat attcccatgg ccaacgacat ttattgatgt tgattatagc
      301 tataatgaat catataacct tatagaagat gtaaagatca ccaaggacac tttggaggag
      361 ttagatgaga tcgccaacaa gagggaggtg gcctactata ggtcagtcat aaacagcctg
      421 aggagcaagg ggtttaaggt tatagttaat ctaaatcact tcacccttcc atattggttg
      481 catgatccca ttgaggctag ggagagggcg ttaactaata agaggaacgg ctgggttaac
      541 ccaagaacag ttatagagtt tgcaaagtat gccgcttaca tagcctataa gtttggagat
      601 atagtggata tgtggagcac gtttaatgag cctatggtgg ttgttgagct tggctaccta
      661 gccccctact ctggcttccc tccaggggtt ctaaatccag aggccgcaaa gctggcgata
      721 cttcacatga taaatgcaca tgctttagct tataggcaga taaagaagtt tgacactgag
      781 aaagctgata aggattctaa agagcctgca gaagttggta taatttacaa caacattgga
      841 gttgcttatc ccaaggatcc gaacgattcc aaggatgtta aggcagcaga aaacgacaac
      901 ttcttccact cagggctgtt cttcgaggcc atacacaaag gaaaacttaa tatagagttt
      961 gacggtgaaa cgtttataga tgccccctat ctaaagggca atgactggat aggggttaat
     1021 tactacacaa gggaagtagt tacgtatcag gaaccaatgt ttccttcaat cccgctgatc
     1081 acctttaagg gagttcaagg atatggctat gcctgcagac ctgggactct gtcaaaggat
     1141 gacagacccg tcagcgacat aggatgggaa ctctatccag aggggatgta cgattcaata
     1201 gttgaagctc acaagtacgg cgttccagtt tacgtgacgg agaacggaat agcggattca
     1261 aaggacatcc taagacctta ctacatagcg agccacataa agatgacaga gaaggccttt
     1321 gaggatgggt atgaagttaa gggctacttc cactgggcat taactgacaa cttcgagtgg
     1381 gctctcgggt ttagaatgcg ctttggcctc tacgaagtca acctaattac aaaggagaga
     1441 attcccaggg agaagagcgt gtcgatattc agagagatag tagccaataa tggtgttacg
     1501 aaaaagattg aagaggaatt gctgagggga tga
//
```

7. Klik “Distance tree of results” Apabila ingin mengetahui phylogenetic tree antar sekuen yang didapatkan. Sebelum melakukan analisis ini, harus dipilih database sekuen yang akan dibandingkan.



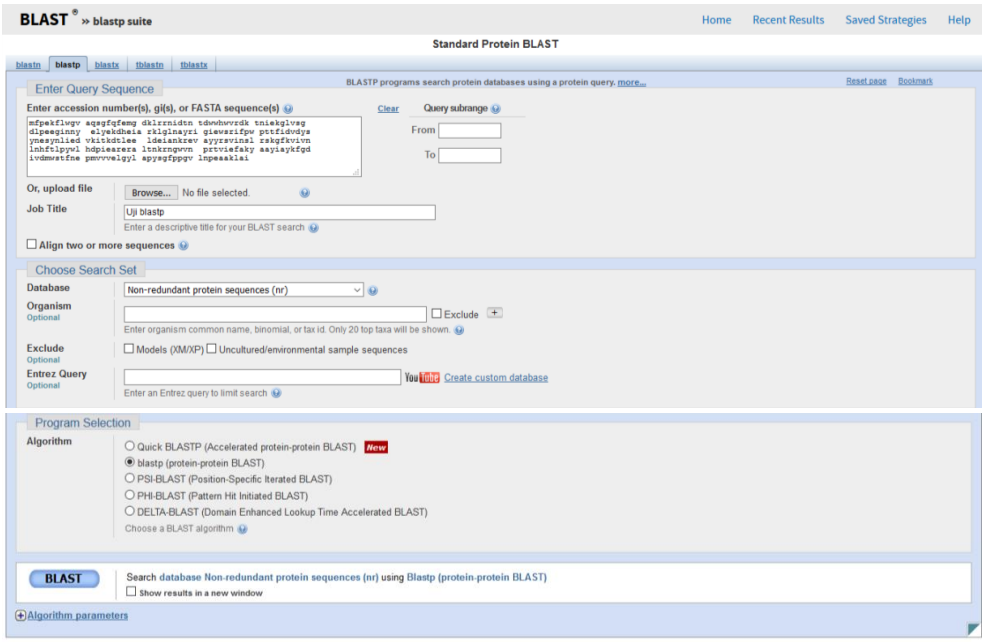
Langkah-langkah analisis menggunakan **blastp**

Blastp dapat digunakan untuk mencari protein homolog dari protein yang kita miliki.

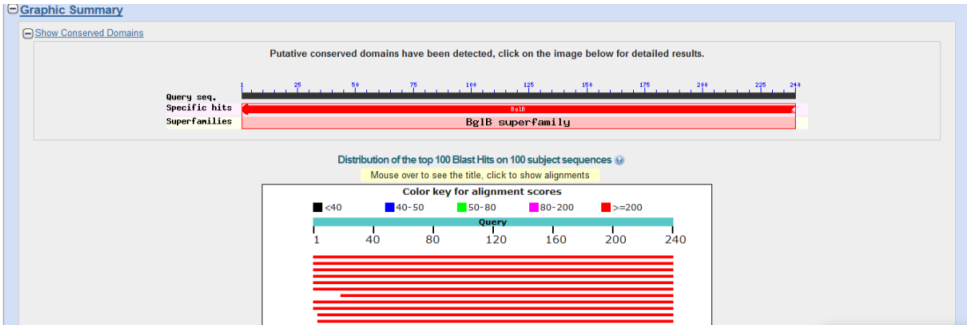
1. Buka situs www.ncbi.nlm.nih.gov

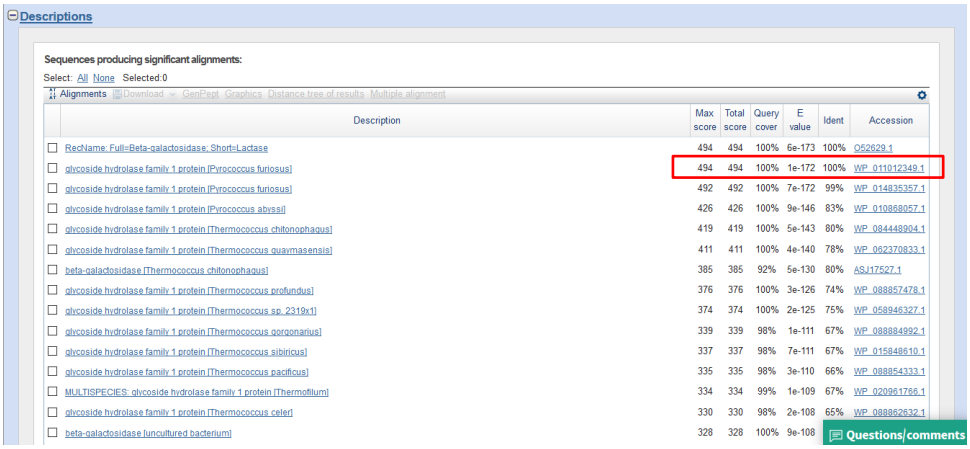
2. Pilih tool ”BLAST”. Untuk mencari protein homolog dari query asam amino gunakan ”protein blast” (blastp)
3. Setelah tampilan muncul, entri sekuen protein (*query*) yang akan dicari; pilih seting pencarian dari database (jika membatasi hanya ingin mencari pada spesies tertentu, ketik nama organisme); pilih program ”blastp”; klik ”BLAST” untuk memulai proses *searching*.

   Pada latihan / contoh digunakan query sekuen protein berikut ini :
   mfpekflwgv aqsgfqfemg dklrrnidtn tdwwhwvrdk tniekglvsg dlpeeginny elyekdheia rklglnayri giewsrifpw pttfidvdys ynesynlied vkitkdtlee ldeiankrev ayyrsvinsl rskgfkvivn lnhftlpywl hdpiearera ltnkrngwvn prtviefaky aayiaykfgd ivdmwstfne pmvvvelgyl apysgfppgv lnpeaaklai



4. Hasil *searching* akan didapat tampilan seperti berikut:





5. Hasil blast akan menghasilkan lebih dari satu sekuen yang bersesuaian. Pilih hasil dengan skor paling tinggi. Dengan meng-klik referensi akan didapat keterangan lebih lanjut tentang protein tersebut.

glycoside hydrolase family 1 protein [Pyrococcus furiosus]

NCBI Reference Sequence: WP_011012349.1

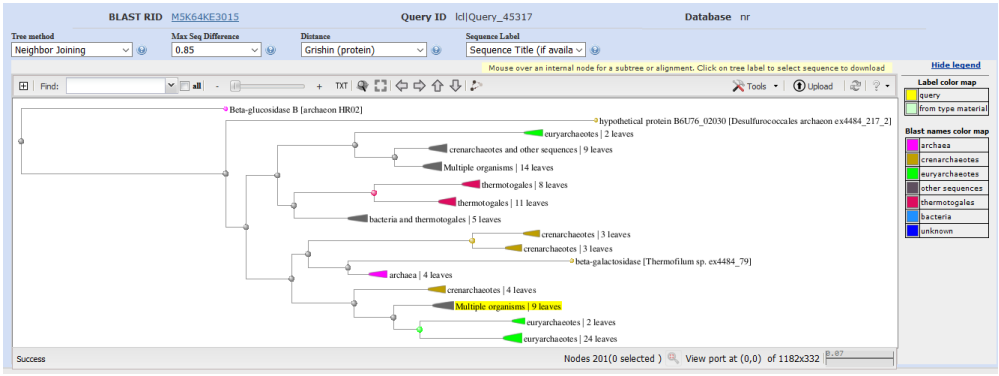
Identical Proteins FASTA Graphics

Go to:

```
LOCUS       WP_011012349              510 aa            linear   BCT 04-JUL-2017
DEFINITION  glycoside hydrolase family 1 protein [Pyrococcus furiosus].
ACCESSION   WP_011012349
VERSION     WP_011012349.1
KEYWORDS    RefSeq.
SOURCE      Pyrococcus furiosus
  ORGANISM  Pyrococcus furiosus
            Archaea; Euryarchaeota; Thermococci; Thermococcales;
            Thermococcaceae; Pyrococcus.
COMMENT     REFSEQ: This record represents a single, non-redundant, protein
            sequence which may be annotated on many different RefSeq genomes
            from the same, or different, species.
            COMPLETENESS: full length.
```

```
FEATURES             Location/Qualifiers
     source          1..510
                     /organism="Pyrococcus furiosus"
                     /db_xref="taxon:2261"
     Protein         1..510
                     /product="glycoside hydrolase family 1 protein"
                     /calculated_mol_wt=58938
     Region          1..504
                     /region_name="BglB"
                     /note="Beta-glucosidase/6-phospho-beta-glucosidase/beta-
                     galactosidase [Carbohydrate transport and metabolism];
                     COG2723"
                     /db_xref="CDD:225343"
ORIGIN
        1 mfpekflwgv aqsgfqfemg dklrrnidtn tdwwhwvrdk tniekglvsg dlpeeginny
       61 elyekdheia rklglnayri giewsrifpw pttfidvdys ynesynlied vkitkdtlee
      121 ldeiankrev ayyrsvinsl rskgfkvivn lnhftlpywl hdpiearera ltnkrngwvn
      181 prtviefaky aayiaykfgd ivdmwstfne pmvvvelgyl apysgfppgv lnpeaaklai
      241 lhminahala yrqikkfdte kadkdskepa evgiiynnig vaypkdpnds kdvkaaendn
      301 ffhsglffea ihkgklnief dgetfidapy lkgndwigvn yytrevvtyq epmfpsipli
      361 tfkgvqqygy acrpgtlskd drpvsdigwe lypegmydsi veahkygvpv yvtengiads
      421 kdilrpyyia shikmiekaf edgyevkgyf hwaltdnfew algfrmrfgl yevnlitker
      481 ipreksvsif reivanngvt kkieeellrg
//
```

6. Klik “Distance tree of results” pada bagian akhir apabila ingin mengetahui phylogenetic tree antar protein yang didapatkan. Sebelum melakukan analisis ini, harus dipilih database protein yang akan dibandingkan.



Bab 3

Analisis Pensejajaran Sekuens

3.1 Pendahuluan

Dalam bioinformatika, Sequence Alignment (SA) adalah suatu cara untuk mencocokkan sequence DNA, RNA atau protein untuk mengetahui tingkat kemiripan. Jika panjang antara sequences yang dibandingkan tidak sama maka dapat diberikan gaps atau yang disebut dengan area di dalam grafik untuk mengisi kekosongan.

Penyejajaran sekuens (***sequence alignment***) dapat diartikan juga proses penyusunan/pengaturan dua atau lebih sekuens sehingga persamaan sekuens-sekuens tersebut tampak nyata. Hasil dari proses tersebut juga disebut sebagai *sequence alignment* atau *alignment* saja. Baris sekuens dalam suatu *alignment* diberi sisipan (umumnya dengan tanda "–") sedemikian rupa sehingga kolom-kolomnya memuat karakter yang identik atau sama di antara sekuens-sekuens tersebut. Berikut adalah contoh *alignment* DNA dari dua sekuens pendek DNA yang berbeda, "ccatcaac" dan "caatgggcaac" (tanda "|" menunjukkan kecocokan atau *match* di antara kedua sekuens).

```
ccat---caac
|  |     |
caatgggcaac
```

Sequence alignment merupakan metode dasar dalam analisis sekuens. Metode ini digunakan untuk mempelajari evolusi sekuens-sekuens dari leluhur yang sama (*common ancestor*). Ketidakcocokan (*mismatch*) dalam *alignment* diasosiasikan dengan proses mutasi, sedangkan kesenjangan (*gap*, tanda "–") diasosiasikan dengan proses insersi atau delesi. *Sequence alignment* memberikan hipotesis atas proses evolusi yang terjadi dalam sekuens-sekuens tersebut. Misalnya, kedua sekuens dalam contoh *alignment* di atas bisa jadi berevolusi dari sekuens yang sama "ccatgggcaac". Dalam kaitannya dengan hal ini, *alignment* juga dapat menunjukkan posisi-posisi yang dipertahankan (*conserved*) selama evolusi dalam sekuens-sekuens protein, yang menunjukkan bahwa posisi-posisi tersebut bisa jadi penting bagi struktur atau fungsi protein tersebut.

Selain itu, *sequence alignment* juga digunakan untuk mencari sekuens yang mirip atau sama dalam basis data sekuens. BLAST adalah salah satu metode *alignment* yang sering digunakan dalam penelusuran basis data sekuens. BLAST menggunakan algoritma heuristik dalam penyusunan *alignment*.

Secara umum, dalam sequences alignment dikenal 2 jenis, yaitu Global Alignment (membandingkan 1 rangkaian sequences secara penuh) dan Local Alignment (hanya mencari subsequences yang mirip) yang akan dijelaskan dibawah ini.

3.2 Jenis – jenis Pensejajaran Sequence

1. Local Alignment

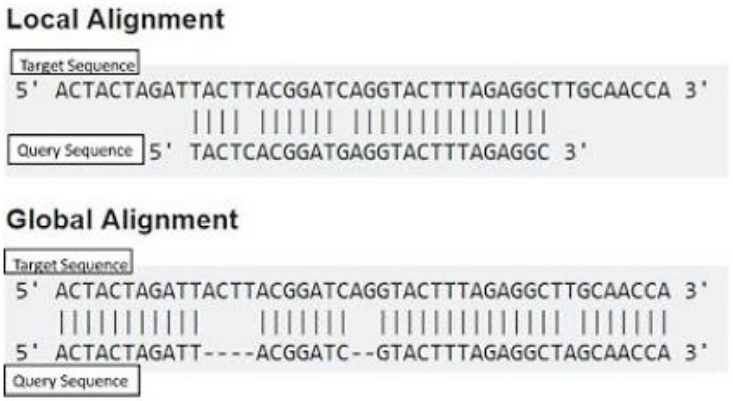
Adalah Urutan yang diduga memiliki kesamaan atau bahkan urutan yang berbeda dapat dibandingkan dengan metode penyelarasan lokal. Ia menemukan daerah lokal dengan tingkat kemiripan yang tinggi.

Perataan lokal meluruskan substring dari urutan kueri ke substring dari urutan target. Substring dapat berupa salah satu atau kedua urutan; jika semua keduanya dimasukkan maka pelurusan lokal juga bersifat global. Perataan lokal didefinisikan dengan memaksimalkan skor penyejajaran, sehingga menghapus kolom dari salah satu ujung akan mengurangi skor, dan menambahkan kolom lebih lanjut di kedua ujung juga akan mengurangi skor. Misalnya, pertimbangkan keselarasan protein global ini.

2. Global Alignment

Adalah Urutan terkait erat yang memiliki panjang yang sama sangat tepat untuk penyelarasan global. Di sini, penyelarasan dilakukan dari awal hingga akhir dari urutan untuk menemukan keselarasan terbaik yang mungkin Sebuah rumus atau serangkaian langkah untuk memecahkan masalah dikembangkan oleh Saul B. Needleman dan Christian D. Wunsch pada tahun 1970, yang merupakan algoritma pemrograman dinamis untuk penyelarasan urutan. Pemrograman dinamis memecahkan masalah asli dengan membagi masalah menjadi masalah sub independen yang lebih kecil. Teknik-teknik ini digunakan dalam berbagai aspek ilmu komputer. Algoritma ini menjelaskan keselarasan urutan global untuk menyelaraskan urutan nukleotida atau protein.

Berikut perbandingan antara local dan global sequence alignment



GLOBAL SEQUENCE ALIGNMENT	LOCAL SEQUENCE ALIGNMENT
Dalam pensejajaran global, dilakukan upaya untuk mensejajarkan seluruh urutan (ujung ke ujung keselarasan).	Menemukan daerah lokal dengan tingkat tertinggi kesamaan antara dua urutan.
Perataan global mengandung semua huruf dari keduanya urutan pertanyaan dan target.	Pensejajaran lokal meluruskan substring pertanyaan urutan ke substring dari urutan target.
Jika dua urutan memiliki kurang lebih sama panjang dan sangat mirip, mereka cocok untuk penyelarasan global.	Setiap dua urutan dapat disejajarkan secara lokal sebagai pensejajaran lokal menemukan barisan-barisan dengan tingkat tinggi cocok tanpa mempertimbangkan penyelarasan sisa bagian urutan.
Cocok untuk menyelaraskan dua yang terkait erat urutan.	Cocok untuk mensejajarkan urutan yang lebih berbeda atau terkait urutan jauh.
Keberpihakan global biasanya dilakukan untuk membandingkan gen homolog seperti membandingkan dua gen dengan fungsi yang sama (pada manusia vs. tikus) atau membandingkan dua protein dengan fungsi serupa.	Digunakan untuk mencari tahu pola-pola yang diawetkan dalam urutan DNA atau domain atau motif yang diawetkan menjadi dua protein.
Teknik pensejajaran global umum adalah Algoritma Needleman-Wunsch	Metode pensejajaran lokal umum adalah Smith–Algoritma Waterman.
Contoh alat pensejajaran Global: ➢ EMBOSS Needle ➢ Needleman-Wunsch Global Align ➢ Nucleotide Sequences (Specialized ➢ BLAST)	Contoh alat pensejajaran Lokal: ➢ BLAST ➢ EMBOSS Water ➢ LALIGN

3.3 Metode –metode Pensejajaran Sequence

Untuk membandingkan 2 sekuens yang paling dikenal ada 3 metode, yaitu

1. Metode Dot-Matrixs

Merupakan metode yang paling mudah. sekuens yang dicocokkan dibuatkan sebuah matrix dimana kolom dan baris mereprentasikan masing sekuens. Jika ada yang sama (match) maka diberi simbol (dot/titik). Sequences yang berkorelasi akan terlihat membentuk garis di diagonal utama. Secara visual dapat dilihat fitur dari sekuens yang dibandingkan (insertions, deletions, repeats, or inverted repeats).

Masalah dari metode ini adanya noise serta sulit mengekstrak fitur posisi kecocokan antara dua sekuens.

2. Metode Dynamic programming

Metode ini dapat diterapkan untuk mencari Global Alignment maupun Local Alignment dari 2 sekuens. Metode yang terkenal adalah Algoritma Needleman-Wunsch (Global) dan Smith-Waterman (Local). Needleman-Wunsch merupakan metode pertama yang menerapkan Dynamic Programming dalam Sequence Alignment sedangkan Smith-Waterman memodifikasi metode dari Needleman-Wunsch untuk mencari Local Alignment. Dynamic Programming sangat optimal untuk mencari kecocokan antara 2 sekuens dan dapat diperluas untuk lebih dari 2 sekuens. Metode ini sangat lambat untuk sekuens yang sangat panjang.

3. Words Method

Dikenal juga sebagai k-tuple methods, merupakan metode *heuristic* dimana tidak menjamin hasilnya optimal tetapi lebih efisien dibandingkan Dynamic Programming. Sering dipakai untuk pencarian dalam database dengan ukuran yang sangat besar (FASTA, BLAST)

Metode-Metode Multiple Sequence Alignment

Dari perbandingan antara 2 sekuens, kemudian berkembang menjadi Multiple Sequence Alignment (MSA), yaitu lebih dari 2 sekuens yang

dibandingkan. Salah satu tujuannya adalah untuk merekonstruksi phylogenetic tree atau untuk mengetahui tingkat kekerabatan spesies.

Untuk MSA juga ada 3 metode yang terkenal, yaitu

1. Dynamic Programming

Secara teori, metode ini dapat mencocokkan banyak sekuens tetapi sangat "mahal" dalam segi kompleksitas dan memori. Sama seperti Dynamic Programming yang dijelaskan untuk 2 sekuens, algoritma yang terkenal adalah Needleman-Wunsch dan Smith-Waterman.

2. Progressive Alignment

Cara kerjanya adalah mencocokkan sekuens yang terdekat kemudian menambahkan sequence berikutnya sehingga terbentuk tree. Initial tree untuk mencocokkan memakai metode yang mirip dengan FASTA. Contoh metode ini adalah Clustal W dan T-Coffee Metode ini sangat cepat, namun ketika sudah terbentuk tree tidak bisa menyisipkan sekuens yang terbaru yang lebih mendekati kemiripannya. Sehingga keakuratannya tergantung dari inisialisasi awal.

3. Metode Iterasi

Metode ini mencoba memperbaiki kelemahan dari metode Progressive yang tergantung dari inisialisasi. Metode iterasi mencoba mengoptimalkan berdasarkan fungsi scoring dan dapat mencocokkan ulang (realigning) subset sequence. Metode ini sangat lambat dan hasilnya tergantung probabilitas karena didasari stokastik. Contohya adalah Hidden Markov Models dan Algoritma Genetik.

3.4 Pensejajaran Sequence Online

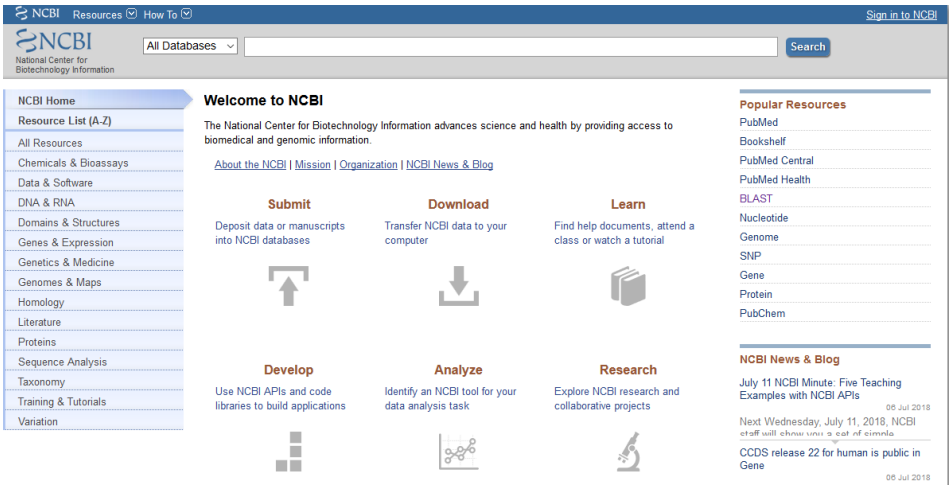
1. Cara mendapatkan data di NCBI

Sekuen yang diperoleh dari hasil penelitian di laboratorium dapat dianalisis dengan data serupa yang telah dipublikasikan sebelumnya di gen bank. Salah satu bentuk analisis yang dapat dilakukan misalnya adalah analisis penyejajaran. Analisis pensejajaran dapat digunakan untuk membandingkan dua sekuen atau lebih. Program yang digunakan untuk

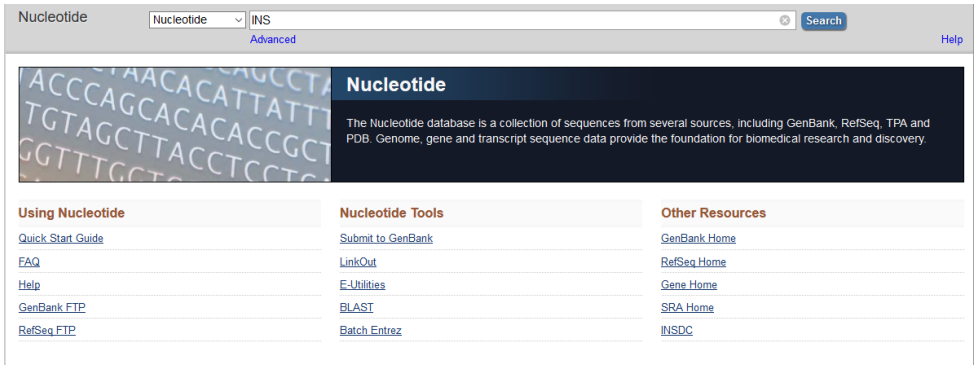
analisis penyejajaran yaitu program BLAST (*Basic Local Allignment Search Tools*). Program ini dapat diakses melalui website National Center for Biotechnology Information at The National Library of Medicine in Washington, DC (http://www.ncbi.nlm.nih.gov/BLAST)

Berikut merupakan langkah-langkah untuk mencari dan mendapatkan data dari *genbank*, misalnya untuk mencari sekuen insulin (INS)

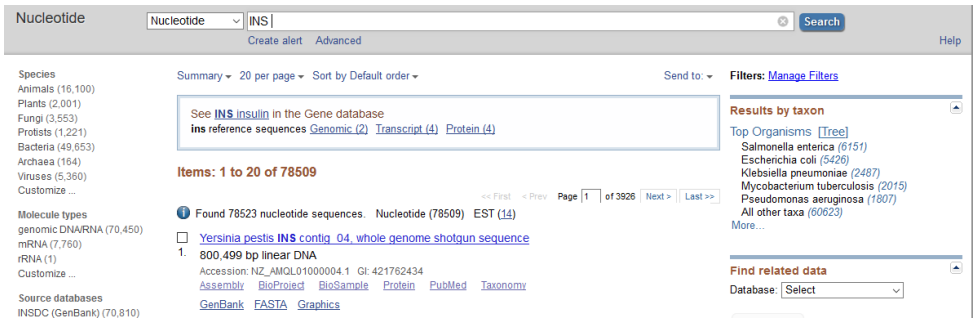
1. Ketikkan **http://www.ncbi.nlm.nih.gov** pada location bar pencarian



2. Pilih preferensi pencarian yang digunakan (pada contoh ini dipilih **nucleotide**) dan ketikkan juga molekul yang ingin dicari sebagai kata kunci pencarian (pada contoh ini diketikkan **INS**) dan klik search

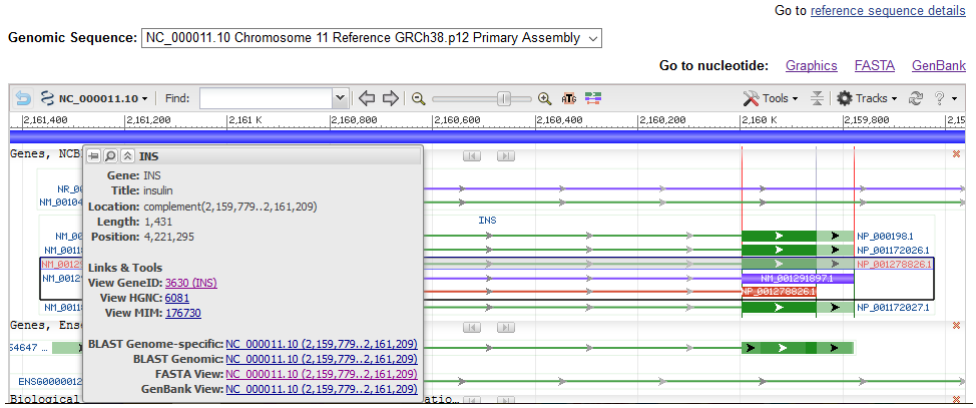


3. Muncul hasil pencarian nucleotide untuk INS insulin. Selanjutnya **klik INS insulin**



4. **Berdasarkan** pulihan tersebut maka akan diperoleh tampilan sebagai berikut

5. **Pada** tampilan tersebut discroll ke bawah maka akan diperoleh tampilan berikut. Terdapat beberapa kode yaitu **NM dan NP**. Kode **NM** menunjukkan kode untuk memperoleh informasi mengenai **nukleotida**, sedangkan **NP** menunjukkan kode untuk memperoleh informasi mengenai **protein**.



6. Di*klik link* FASTA untuk memperoleh sekuen nukleotida dari INS dalam bentuk FASTA

Homo sapiens chromosome 11, GRCh38.p12 Primary Assembly

NCBI Reference Sequence: NC_000011.10

GenBank Graphics

```
>NC_000011.10:c2161209-2159779 Homo sapiens chromosome 11, GRCh38.p12 Primary
Assembly
AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGGTCTGTTCCAAGGGCCTTTGCGTCAGGT
GGGCTCAGGATTCCAGGGTGGCTGGACCCCAGGCCCCAGCTCTGCAGCAGGGAGGACGTGGCTGGGCTCG
TGAAGCATGTGGGGGTGAGCCCAGGGGCCCCAAGGCAGGGCACCTGGCCTTCAGCCTGCCTCAGCCCTGC
CTGTCTCCCAGATCACTGTCCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTG
GCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAG
CTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCT
GCAGGGTGAGCCAACTGCCCATTGCTGCCCCTGGCCGCCCCCAGCCACCCCCTGCTCCTGGCGCTCCCAC
CCAGCATGGGCAGAAGGGGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAAG
TTCTCTTGGTCACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTGAGGACGGTG
TTGGCTTCGGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCCACTCGCCCCTCAAACAAA
TGCCCCGCAGCCCATTTCTCCACCCTCATTTGATGACCGCAGATTCAAGTGTTTTGTTAAGTAAAGTCCT
GGGTGACCTGGGGTCACAGGGTGCCCCACGCTGCCTGCCTCTGGGCGAACACCCCATCACGCCCGGAGGA
GGGCGTGGCTGCCTGCCTGAGTGGGCCAGACCCCTGTCGCCAGGCCTCACGGCAGCTCCATAGTCAGGAG
ATGGGGAAGATGCTGGGGACAGGCCCTGGGGAGAAGTACTGGGATCACCTGTTCAGGCTCCCACTGTGAC
GCTGCCCCGGGGCGGGGGAAGGAGGTGGGACATGTGGGCGTTGGGGCCTGTAGGTCCACACCCAGTGTGG
GTGACCCTCCCTCTAACCTGGGTCCAGCCCGGCTGGAGATGGGTGGGAGTGCGACCTAGGGCTGGCGGGC
AGGCGGGCACTGTGTCTCCCTGACTGTGTCCTCCTGTGTCCCTCTGCCTCGCCGCTGTTCCGGAACCTGC
TCTGCGCGGCACGTCCTGGCAGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGC
CCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCTGCTCCCT
CTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCACACCCGCCGCCTCCTGCACC
GAGAGAGATGGAATAAAGCCCTTGAACCAGC
```

7. Apabila diklik kode **NM** dan **Gen Bank**, maka akan diperoleh informasi mengenai sekuen nukleotida, sebagai berikut

Homo sapiens chromosome 11, GRCh38.p12 Primary Assembly

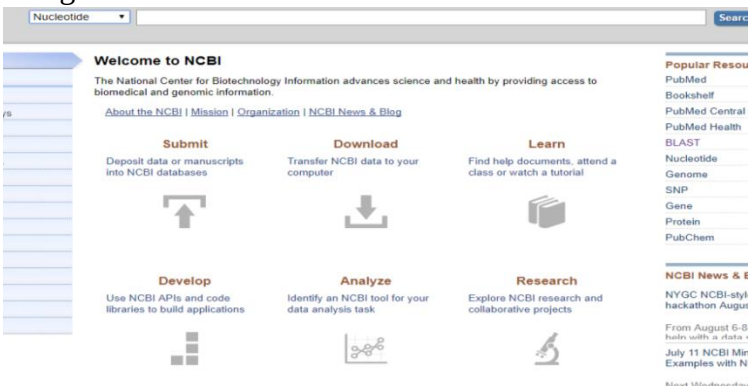
NCBI Reference Sequence: NC_000011.10

FASTA Graphics

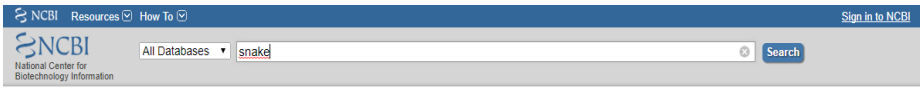
```
LOCUS       NC_000011              1431 bp    DNA     linear   CON 26-MAR-2018
DEFINITION  Homo sapiens chromosome 11, GRCh38.p12 Primary Assembly.
ACCESSION   NC_000011 REGION: complement(2159779..2161209)
VERSION     NC_000011.10
DBLINK      BioProject: PRJNA168
            Assembly: GCF_000001405.38
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1431)
  AUTHORS   Taylor,T.D., Noguchi,H., Totoki,Y., Toyoda,A., Kuroki,Y., Dewar,K.,
            Lloyd,C., Itoh,T., Takeda,T., Kim,D.W., She,X., Barlow,K.F.,
            Bloom,T., Bruford,E., Chang,J.L., Cuomo,C.A., Eichler,E.,
            FitzGerald,M.G., Jaffe,D.B., LaButti,K., Nicol,R., Park,H.S.,
            Seaman,C., Sougnez,C., Yang,X., Zimmer,A.R., Zody,M.C.,
            Birren,B.W., Nusbaum,C., Fujiyama,A., Hattori,M., Rogers,J.,
            Lander,E.S. and Sakaki,Y.
  TITLE     Human chromosome 11 DNA sequence and analysis including novel gene
            identification
  JOURNAL   Nature 440 (7083), 497-500 (2006)
```

2. Persejajaran Sequence Online

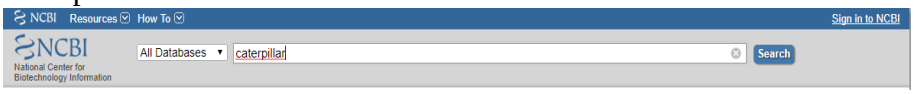
1. Ketik NCBI pada search google dan masuk website NCBI
2. Carilah jenis makhluk hidup dengan mempunyai kemiripan pada keduanya untuk di sejajarkan
3. Misalnya membandingkan snake dengan caterpillar.
4. Pada NCBI kita bisa isi all database sesuai yang kita ingin cari, misalnya kita ganti nucleotide.



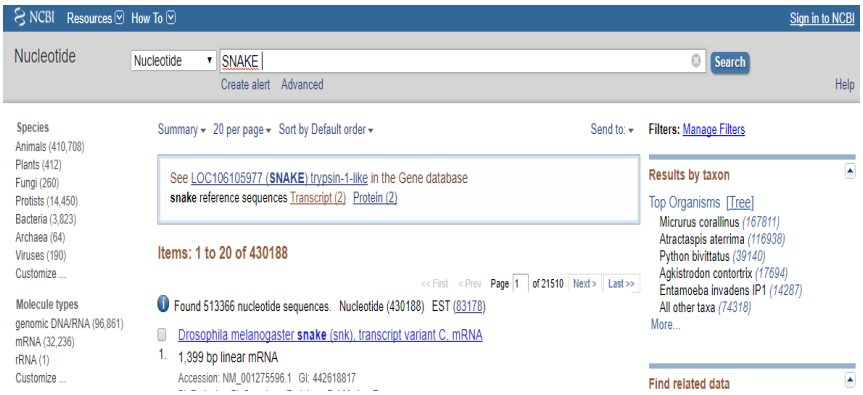
5. Kemudian sebelah kotak all database kita isi sesuai jenis makhluk hidup yang kita ingin sejajarkan.
   - Snake



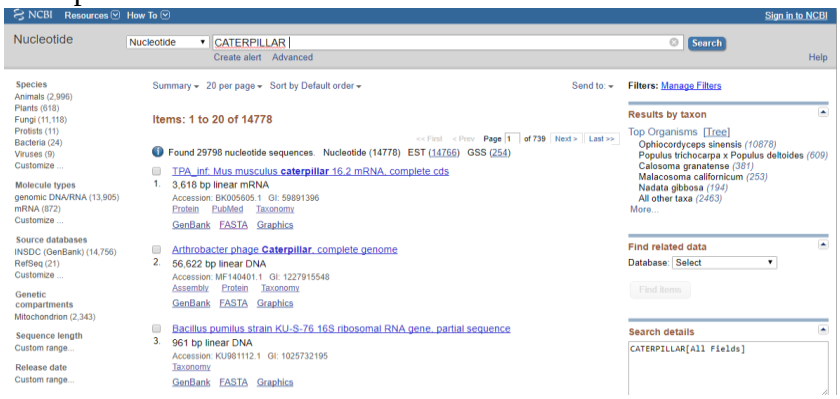
   - Caterpillar



6. Setelah kita search kita ambil salah satu data dari beberapa yang muncul.
   - Snake



- Caterpillar



7. Kemudian klik FASTA
   - Snake

   - Caterpillar

8. Copy kode FASTA
   - Snake

GCTTCTTATATGAAATGTTTGAAAGAAAGTAAATAATGAAATCACAGAACTCACAAA
TAGAATGATAATACTTTGGTCTCTAATCGTGCATCTTCAATTGACATGCTTGCATTT
AATCCTGCAGACTCCAAATCTTGAGGCTCTCGATGCCCTTGAAATAATCAACTACCAG
ACCACCAAGTACACAATCCCCGAAGTTTGGAAAGAGCAGCCAGTGGCGACCATTGGC
GAGGACGTGGATGATCAGGACACCGAAGATGAGGAAAGCTATCTGAAGTTCGGCGAC
GATGCGGAGGTTAGGACTTCGGTTTCGGAAGGACTTCACGAAGGTGCCTTCTGCCGG
CGGAGCTTCGATGGCCGGAGTGGATACTGCATCCTGGCCTATCAGTGCCTCCACGTCA
TTCGAGAGTATCGGGTGCATGGCACCCGCATCGACATCTGCACCCATCGCAACAATGT
CCCTGTCATCTGTTGTCCATTGGCGGACAAGCATGTGCTGGCTCAGCGCATAAGTGCC
ACCAAATGTCAGGAGTACAATGCTGCCGCCAGAAGGCTTCACTTGACGGACACCGGA
CGCACATTCTCCGGGAAGCAGTGTGTGCCCAGTGTTCCCCTGATAGTCGGTGGAACCC
CCACTCGACACGGACTCTTCCCTCACATGGCTGCCTTAGGATGGACGCAGGGTAGTGG
CTCCAAGGATCAAGATATAAAATGGGGCTGTGGAGGCGCCCTGGTTAGCGAACTGTA
TGTCCTGACCGCTGCCCACTGTGCCACCTCTGCAAACCACCGGACATGGTTCGCTTGG
GCGCCCGCCAGTTGAACGAGACCAGCGCGACCCAGCAGGACATCAAGATCCTCATCAT
CGTGCTGCATCCGAAGTACAGATCCTCGGCATATTACCACGATATTG
CCCTGCTCAAGCTGACCAGAAGGGTCAAGTTCTCGGAGCAGGTGCGTCCTGCTTGCCT
GTGGCAACTGCCGGAGCTCCAGATACCCACTGTGGTGGCCGCCGGTTGGGGACGCACC

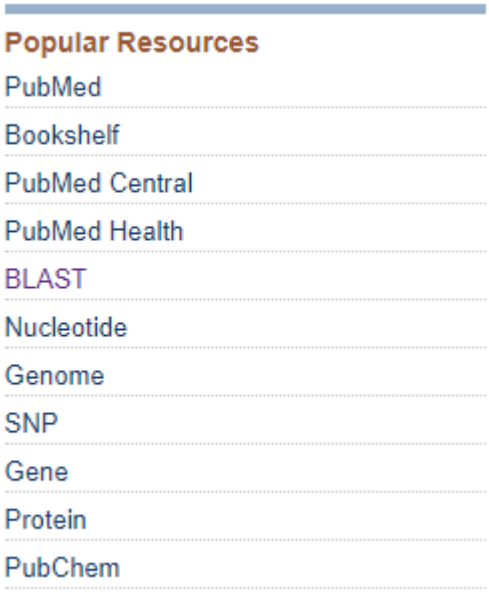
GAGTTCCTGGGCGCCAAATCGAATGCCCTGCGCCAGGTGGACCTGGACGTAGTCCCAC
AAATGACCTGCAAGCAGATCTATCGCAAGGAGCGACGTCTGCCCAGGGGAATCATCG
AGGGGCAGTTCTGTGCGGGATATTTGCCAGGCGGCAGGGACACCTGCCAGGGTGACT
CCGGCGGTCCCATTCATGCCCTGCTGCCGGAATACAACTGCGTGGCCTTCGTGGTGGG
CATCACCTCGTTTGGAAAATTCTGTGCGGCTCCCAATGCCCCAGGAGTTTACACCAGG
CTATATAGCTACCTGGATTGGATTGAGAAGATTGCCTTCAAGCAGCACTAGTTTCAT
TTTATTTTATTTATTAAATATGCTTTTT

- Caterpillar

ATGGAGCTCAGCCGAGTGGGAGACAACATAGGTTCCCCAGGGTCAGTCCTGGCTCTG
TATTCACAACTTCTGGCTGCAAACACAGACTCCACGAGGAAGCAAGAGGTGTGGACA
GACAGAGAGACATGCCTGGCCTACAGTGTTGGCTCCCCAGCTGAGCAGGTGAAAGCC
CTTGTGGATCTGCTGGCTGGGAAGGGCAGTCAGCTGCTACAAGTCCGGGACAAAATG
CCAGACTCCCCACTAGGATCCCAGAGCAATGAGTCAAGGATACCGAAGCACTCTGAG
GCTCTGCTGAGCAGGGTGGGAAATGACCCAGAACTGGGCAGCCCCTCACACCGGCTGG
CCAGCCTCATGCTGGTCGAGGGCCTGACAGACCTGCAGCTAAAGGAGCATGACTTCAC
ACAGGTGGAGGCCACGCGTGGGGTCTGGCACCCTGCCAGAGTTATCACCCTGGACAGG
CTCTTCCTGCCTCTGTCCCGGGTATCCATCCCACCTCGAGTCTCTCTCACCATTGGAG
TGGCTGGTGTGGGCAAGACCACGCTAGTGAGGCATTTTGTTCATTGCTGGGCCAGAG
GACAGGTGGGCAAGGGCTTCTCACGGGTTCTGCCCTTGACCTTTCGGGATCTCAACAC
CTATGAGAAACTGTCTGCAGACAGACTCATCCAATCCATCTTCTCAAGCATTGGGGA
AGCTAGTCTGGTGGCCACAGCCCCAGACAGAGTCCTCCTGGTCCTGGATGGCTTGGAT
GAGTGTAAGACACCCCTGGAATTCTCCAATACCATGGCCTGCTCAGACCCAAAGAAG
GAGATCCAGGTAGACCACCTGATCACTAACATCATCCGAGGCAACCTCTTTCCAGAA
ATTTCTGTCTGGATCACCTCCCGGCCCAGTGCTGCTGGTCAGATCCCTG
GGGGCCTAGTGGACCGGATGACTGAGATTCGGGGCCTTACTGAGGAAGAGATCAAAG
TGTGTCTGGAGCAGATGTTTCCTGAGGAGCAGAACCTCTTAGGTCAGGTCCTTAGTC
AAGTGCAGGCCAACAGGGCTCTGTATCTGATGTGCACTGTACCAGCCTTTTGTAGGC
TCACGGGGCTGGCTCTGGGTCACTTGTATCGCACCAGGCTGGCCGTCCAAGACATAGA
GCTGCCATTGCCTCAGACCCTGTGTGAGCTCTACTCTTGGTACTTTAGGATGGCTCTT
GGTGGGGAGGGCCAGGATAAGGAAAAGGTAAGTCCTAGGATCAAGCAGGTGACCCAG
GGAGCTCGCAAAATGGTGGGGACATTGGGCCGCCTGGCCTTCCATGGGCTGGTCAAG
AAGAAATACGTGTTTTATGAACAAGACATGAAGGCATTTGGAGTGGACCTCGCTCTG
TTGCAGAACACTCTGTGCAGCTGTCTCCTGCAGCGGGAAGAGACCCTGGCCTCCTCTG
TAGCTTACTGCTTCATTCACCTGTCTCTGCAAGAATTTGTGGCAGCTACATATTACTA
TAGTGCATCCAAGAGGGCCATCTTTGACCTCTTCACCGAGAGTGGCATGTCCTGGCCC
AGACTGGGTTTCCTCGCCCATTTCAGGTGTGCAGCCCAGCGGGCCACACAAGCTAAGG
ATGGAAGGCTGGATGTGTTTCTGCGCTTCCTCTCTGGCCTCTTGTCCCCAAGGGTCAA
TACTCTGCTGGCCGGCTCCCTGTTGTCCCAAGGCGAGCATCAGAGCTACCGGGACCAG
GTGGCTGAGGTCCTACAAGGCTTCCTTCATCCTGACGCTGCAGTCTGTGCACGTGCCA
TCAATGTCTTGTACTGCCTAAGTGAGCTGCGGCACACAGAACTGGC
CTGCAGTGTGGAGGAGGCCATGCGGAGTGGGACCTTGGCTGGGATGACCAGCCCCTC
ACACCGCACTGCTCTGGCCTACCTCCTGCAGATGTCTGACATCTGCTCCCCAGAGGCT
GACTTCTCCCTGTGTCTCAGCCAGCATGTCCTCCAGAGCCTGCTGCCCCAGCTGCTCT
ATTGTCAAAGCCTCAGGCTGGACAACAACCAGTTCCAGGACCCTGTGATGGAGTTGC
TGGGCAGCGTGCTGAGTGGGAAGGACTGTCGCATTCGAAAGATCAGCCTGGCTGAGA
ATCAGATTGGTAACAAAGGAGCCAAAGCCCTGGCCAGATCCCTCCTGGTTAACAGAA
GCCTCATCACACTGGACCTCCGGAGTAACAGCATTGGACCACCGGGGGCTAAGGCTTT
GGCCGATGCTCTGAAGATAAACCGAACGCTAACTTCTCTAAGCCTCCAAAGCAACGT
GATCAAGGATGACGGTGTCATGTGCGTGGCTGAGGCCCTGGTCTCCAACCAGACCATC
TCCATGCTACAGCTACAGAAGAACTTAATTGGGCTCATAGGAGCCCAGCAGATGGCA

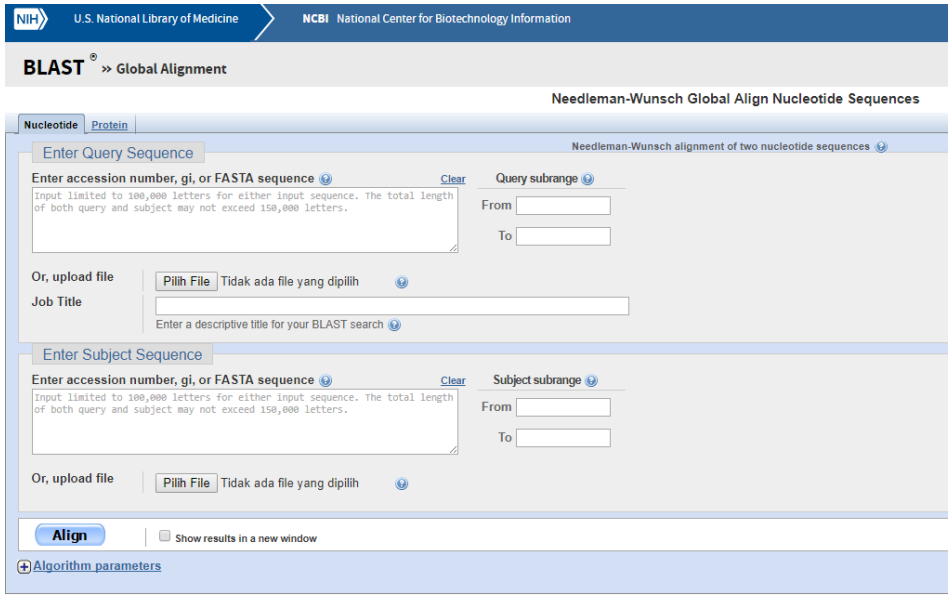
GATGCCCTGAAGCAGAACAGGAGCCTGAAAGCACTCATGTTTTCCAGTAATACCATT
GGCGACAGAGGTGCCATAGCCCTGGCTGAGGCCCTGAAGGTGAACCAGATCCTGGAG
AACTTAGACCTACAGAGCAATTCCATCAGTGACATGGGAGTGACGGTGCTGATGCGA
GCCCTCTGCAGTAACCAGACACTCTCCAGTCTCAACTTACAGGAGAATGCCATAGGGG
ATGAAGGAGCTTCCTCAGTGGCTGGCGCACTGAAGGTGAACACAACCCTCATTGCTC
TCTACTTACGAGGAAACGACGTTGGGGCAGCTGGAGCCAAGGCCTTGGCA
AATGCTTTAAAGTTAAACTCCAGTCTCCGAAGACTCAATCTCCAGGAGAACTCACTG
GGGATGGATGGGGCCATATTTGTTGCCTCTGCACTGTCTGAGAACCACGGGTCTCCAC
ATGACCCAACTCAAAAGAACATAAGACAAGATAATGACAGTGTTCCTGCCCTGCACA
CTCACATGGCTGGCACCATCAGTCTCTGCAGTGTTCCTGCCCTGCACACTCACACGGA
TGTTACCATCAGTCTCTGTAGTGTTCCTGCCCTGCACACTCCCATGGCTGGCACCATG
AGTCTCTGCAGTGTTCCTGCCCTGCACACTCACATGGCTGGCACCATCAGTCTCTGTA
GTGTTCCTGCCCTGCACACTCCCATGGCTGGCACCATCAGTCTCTGCAGTGTTCCTGC
CCTGCACACTCACACGGTTGGCACCATCAGTCTCTGCAGTGTTCCTGCCCTGCACACT
CACATGGATGTCACCATCAGTCTCTGTAGTGTTCCTGCCCTGCACACTCCCATGGCTG
GCACCATCAGTCTCTGTAGTGTTCCTGCCCTGCACACTCCCATGGCTGGCACCATCAG
TCTCTGCAGTGTTCCTGCCCTGCACACTCACACGGTTGGCACCATCAGTCTCTGCAGT
GTTCCTGCCCTGCACACTCCCATGGCTGGCACCATCAGTCTCTGCAGTGTTCCTGCCC
TGCACACTCCCATGGCTGGCACCATCAGTCTCTGTAGTGTTCCTGCCCTGCACACTCC
CATGGCTGGCACCATCAGTCTCTGTAGTGTTCCTGCCCTGCACACTCCCATGGCTGGC
ACCATCAGTCTCTGTAGTGTTCCTGCCCTGCACACTCATACAGCTGGCACCAGTCTCT
GCAGACTTCAAGCCCACTAA

9. Pilih BLAST pada popular resources

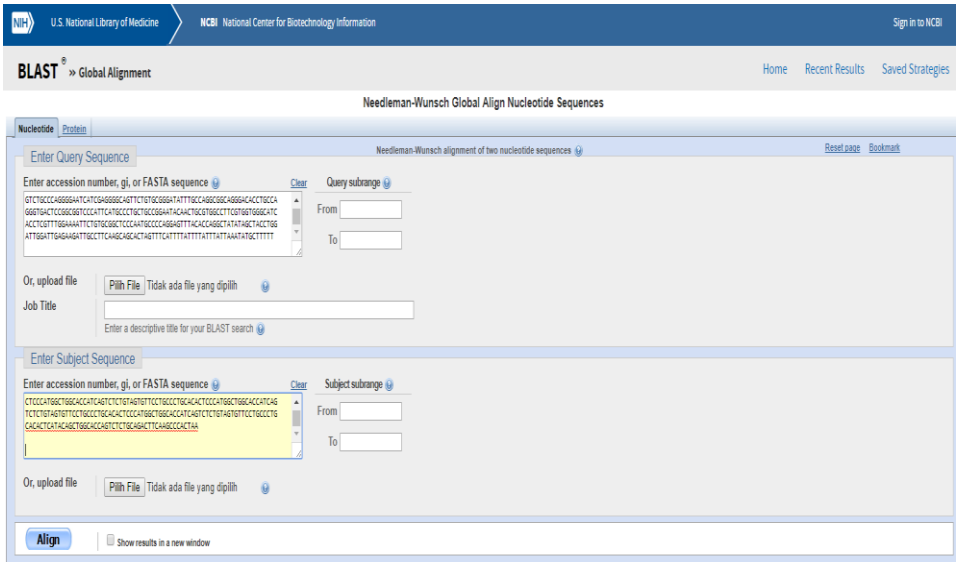


10. Akan tampil menu BLAST kemudian pilih Global Align pada Specialized Seraches seperti berikut

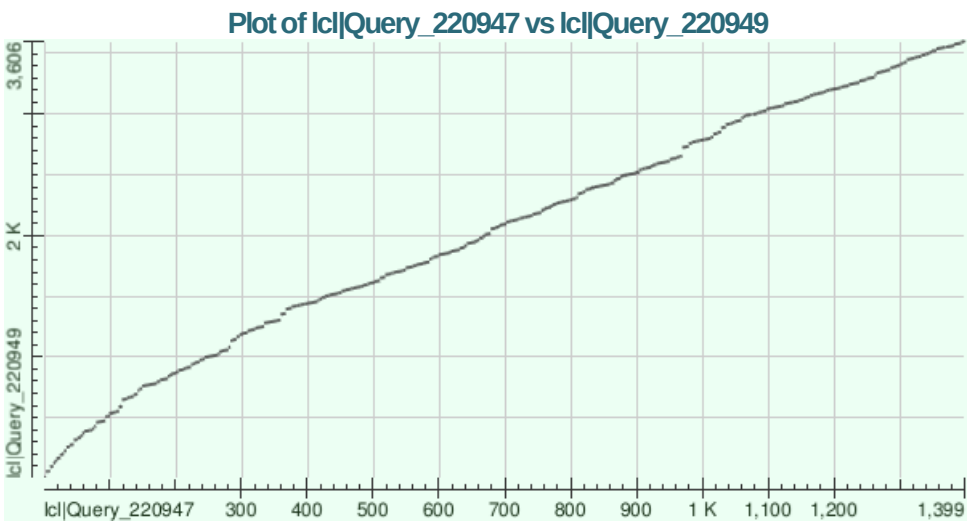
11. Masukkan data kode fasta

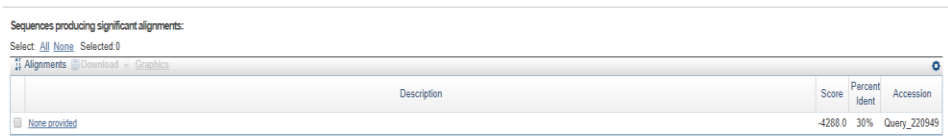


12. Setelah terisi kode fasta kemudian klik **Align**



13. Hasil pensejajaran kode FASTA Snake dan Caterpillar dengan menggunakan BLAST.





Pada gambar tersebut terdapat hasil pensejajaran antara kode FASTA snake dan caterpillar menghasilkan kemiripan 30%.

Bab 4

Mutasi

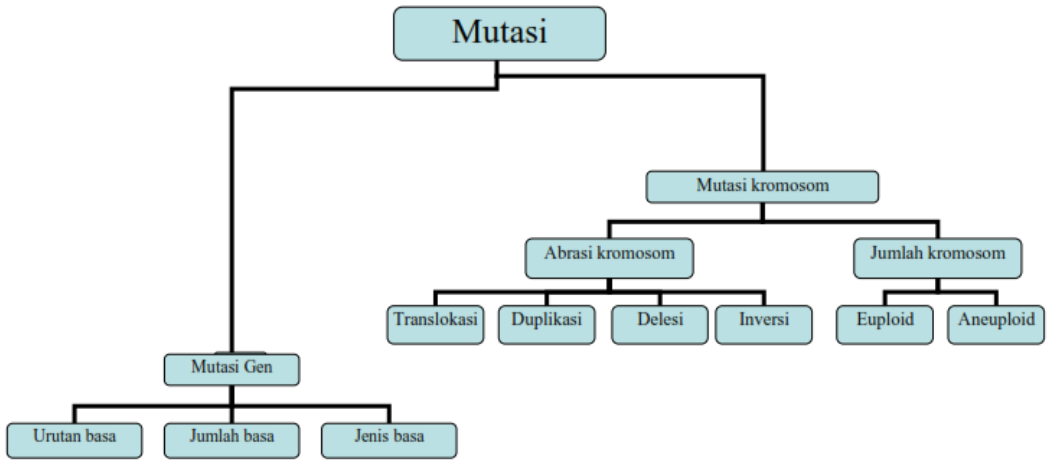
4.1 Definisi Mutasi

Mutasi merupakan perubahan urutan basa nukleotika pada DNA maupun kromosom. Tahun 1901, Hugo De Vries memperkenalkan isilah "mutasi" yang didefinisikan sebagai perubahan informasi pada gen. Selanjutnya, H.J. Muller melihat bahwa peristiwa mutasi spontan dapat dipicu dari paparan sinar rontgen pada *Drosophilla*. Penelitian tentang mutasi terus berkembang, terkait jenis mutasi hingga teknik-teknik untuk menganalisis terjadinya mutasi. Mutasi berasal dari kata Mutatus (bahasa latin) yang artinya adalah perubahan. mutasi didefenisikan sebagai perubahan materi genetic (DNA) yang dapat diwariskan secara genetis keketurunannya. Istilah mutasi petama kali digunakan oleh Hugo de Vries, untuk mengemukakan adanya perubahan fenotipe yang mendadak pada bunga Oenothera lamarckiana dan bersifat menurun. Ternyata perubahan tersebut terjadi karena adanya penyimpangan dari kromosomnya. Seth wright juga melaporkan peristiwa mutasi pada domba jenis Ancon yang berkaki pendek dan bersifat menurun. Penelitian ilmiah tentang mutasi dilakukan pula oleh Morgan (1910) dengan menggunakan Drosophila melanogaster (lalat buah). Akhirnya murid Morgan yang bernama Herman Yoseph Muller berhasil dalam percobaannya terhadap lalat buah, yaitu menemukan mutasi buatan dengan menggunakan sinar X (Anonim, 2009).

Mutasi adalah perubahan pada materi genetik suatu makhluk yang terjadi secara tiba-tiba, acak, dan merupakan dasar bagi sumber variasi organisme hidup yang bersifat terwariskan (heritable). Mutasi juga dapat diartikan sebagai perubahan struktural atau komposisi genom suatu jasad yang dapat terjadi karena faktor luar(mutagen) atau karena kesalahan replikasi. Peristiwa terjadinya mutasi disebut mutagenesis. Makhluk hidup yang mengalami mutasi disebut mutan dan factor penyebab mutasi disebut mutagen (mutagenic agent). Perubahan urutan nukleotida yang menyebabkan protein yang dihasilkan tidak dapat berfungsi baik dalam sel dan sel tidak

mampu mentolerir inaktifnya protein tersebut, maka akan menyebabkan kematian (lethal mutation). Mutasi dapat mempengaruhi DNA maupun kromosom. DNA dapat dipengaruhi pada saat sintesis

DNA (replikasi). Pada saat tersebut factor mutagenic mempengarugi pasangan basa nukleutida sehingga tidak berpasangan dengan basa nukleutida yang seharusnya (mismatch). Misalnya triplet DNA cetakan adalah TTA. Namun karena adanya mutagen menyebabkan DNA polymerase memasangkan A dengan C, bukan dengan T. Untuk lebih jelasnya mekanisme mutasi dapat dilihat pada gambar di bawah ini:



4.2 Jenis mutasi

1. Mutasi Kromosom (*Gross Mutation*)

Mutasi kromosom merupakan perubahan urutan sekuens gen karena terjadi perubahan pada level kromosom. Mutasi kromosom berdampak pada berubahnya struktur dan/atau jumlah kromosom dalam individu. Perubahan struktur kromosom dapat terjadi abrasi kromosom yang melalui proses ***delesi, inversi***, ***duplikasi***, dan ***translokasi***.

1) **Aberasi kromosom** adalah perubahan yang terjadi pada jumlah atau susunan **kromosom** dalam sel yang diakibatkan adanya kehilangan, pengaturan kembali bahan genetik ataupun duplikasi. **Translokasi**

ialah mutasi yang mengalami pertukaran segmen kromosom ke kromosom non homolog. Macam-macam translokasi antara lain sebagai berikut.

1. Translokasi homozigot (resiprok) Translokasi homo zigot ialah translokasi yang mengalami pertukaran segmen kedua kromosom homolog dengan segmen kedua kromosom non homolog.
2. Translokasi heterozigot (non resiprok) Translokasi heterozigot ialah translokasi yang hanya mengalami pertukaran satu segmen kromosom ke satu segmen kromosom nonhomolog.
3. Translokasi Robertson Translokasi Robertson ialah translokasi yang terjadi karena penggabungan dua kromosom akrosentrik menjadi satu kromosom metasentrik, maka disebut juga fusion (penggabungan). Translokasi terjadi apabila dua benang kromosom patah setelah terkena energi radiasi, kemudian patahan benang kromosom bergabung kembali dengan cara baru. Patahan kromosom yang satu berpindah atau bertukar pada kromosom yang lain sehingga terbentuk kromosom baru yang berbeda dengan kromosom aslinya. Translokasi dapat terjadi baik di dalam satu kromosom (intrachromosome) maupun antar kromosom (interchromosome). Translokasi sering mengarah pada ketidakseimbangan gamet sehingga dapat menyebabkan kemandulan (sterility) karena terbentuknya chromatids dengan duplikasi dan penghapusan. Alhasil, pemasangan dan pemisahan gamet jadi tidak teratur sehingga kondisi ini menyebabkan terbentuknya tanaman aneuploidi. Translokasi dilaporkan telah terjadi pada tanaman Aegilops umbellulata dan Triticum aestivum yang menghasilkan mutan tanaman tahan penyakit.
4. Inversi ialah mutasi yang mengalami perubahan letak gen-gen, karena selama meiosis kromosom terpilin dan terjadi kiasma. Inversi terjadi karena kromosom patah dua kali secara simultan setelah terkena energi radiasi dan segmen yang patah tersebut berotasi 180^{o} dan menyatu kembali. Kejadian bila centromere berada pada bagian kromosom yang terinversi disebut pericentric, sedangkan bila centromere berada di luar kromosom yang terinversi disebut paracentric. Inversi pericentric berhubungan dengan duplikasi atau penghapusan chromatid yang dapat menyebabkan

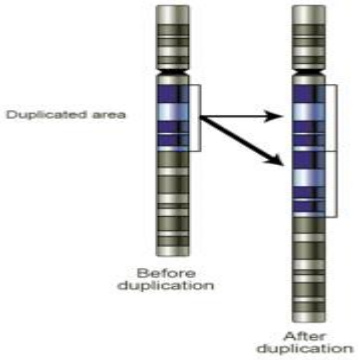
aborsi gamet atau pengurangan frequensi rekombinasi gamet. Perubahan ini akan ditandai dengan adanya aborsi tepung sari atau biji tanaman, seperti dilaporkan terjadi pada tanaman jagung dan barley. Inversi dapat terjadi secara spontan atau diinduksi dengan bahan mutagen, dan dilaporkan bahwa sterilitas biji tanaman heterosigot dijumpai lebih rendah pada kejadian inversi daripada translokasi.

Macam-macam inversi antara lain sebagai berikut.

a. Inversi parasentrik; teriadi pada kromosom yang tidak bersentromer.
b. lnversi perisentrik; terjadi pada kromosom yang bersentromer.

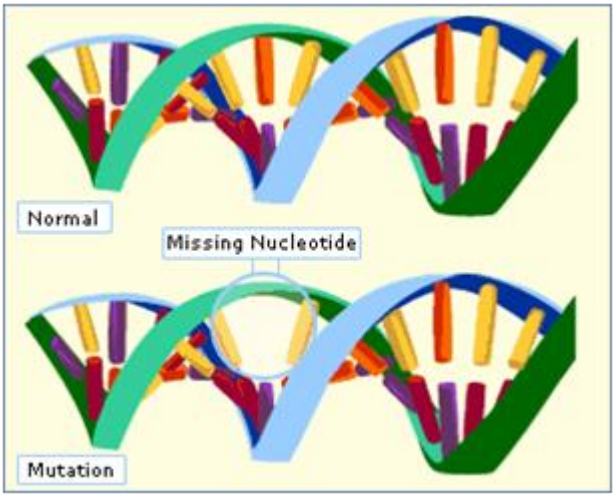
5. Isokromosom ialah mutasi kromosom yang terjadi pada waktu menduplikasikan diri, pembelahan sentromernya mengalami perubahan arah pembelahan sehingga terbentuklah dua kromosom yang masing – masing berlengan identik (sama). Dilihat dari pembelahan sentromer maka isokromosom disebut juga fision, jadi peristiwanya berlawanan dengan translokasi Robertson (fusion) yang mengalami penggabungan.
6. Katenasi ialah mutasi kromosom yang terjadi pada dua kromosom non homolog yang pada waktu membelah menjadi empat kromosom, salinq bertemu ujung-ujungnya sehingga membentuk lingkaran.

Duplikasi gen ataupun **duplikasi kromosom** atau **amplifikasi gen** merupakan kejadian bergandanya (duplikasi) suatu daerah bagian DNA yang mengandung gen. Ia dapat terjadi sebagai kesalahan pada rekombinasi homolog, kejadian retrotransposisi, ataupun duplikasi keseluruhan kromosom. Kopi kedua dari gen ini seringkali terbebas dari tekanan seleksi, yakni bahwa mutasi ini tidak memiliki efek merugikan pada organisme inang. Oleh karenanya, gen ini bermutasi lebih cepat dari generasi ke generasi organisme. Duplikasi merupakan lawan dari delesi. Duplikasi terjadi akibat dari suatu kejadian yang disebut sebagai pindah silang. Ini terjadi semasa meiosis antara kromosom homolog yang salah jajar. Peluang hal ini terjadi adalah berupa fungsi derajat perkongsian elemen berulang antara dua kromosom. Produk rekombinasi adalah duplikasi pada daerah pertukaran dan delesi timbalbalik.



Sebuah skema bagian kromosom sebelum dan sesudah kejadian duplikasi

Delesi, yaitu pengurangan satu atau lebih pasangan nukleotida pada suatu gen. seperti tampak pada gambar di bawah ini:



Inversi ialah mutasi yang mengalami perubahan letak gen-gen, karena selama meiosis kromosom terpilin dan terjadi kiasma. Inversi terjadi karena kromosom patah dua kali secara simultan setelah terkena energi radiasi dan segmen yang patah tersebut berotasi 180^{o} dan menyatu kembali. Kejadian bila centromere berada pada bagian kromosom yang terinversi disebut pericentric, sedangkan bila centromere berada di luar kromosom yang terinversi disebut paracentric. Inversi pericentric berhubungan dengan duplikasi atau penghapusan chromatid yang dapat menyebabkan aborsi gamet atau pengurangan frequensi rekombinasi gamet. Perubahan ini akan ditandai dengan adanya aborsi tepung sari atau biji tanaman, seperti dilaporkan terjadi pada tanaman jagung dan

barley. Inversi dapat terjadi secara spontan atau diinduksi dengan bahan mutagen, dan dilaporkan bahwa sterilitas biji tanaman heterosigot dijumpai lebih rendah pada kejadian inversi daripada translokasi.
Macam-macam inversi antara lain sebagai berikut.
a) Inversi parasentrik; teriadi pada kromosom yang tidak bersentromer.
b) lnversi perisentrik; terjadi pada kromosom yang bersentromer.

Perubahan jumlah kromosom dalam makhluk hidup dapat dibedakan menjadi
dua, yaitu:

a. Euploidi

Euploidi merupakan mutasi kromosom yang menyebabkan perubahan set kromosom dalam tubuh individu. Individu normal adalah individu diploid (2n). Manusia normal memiliki 46 kromosom (2n=46). Perubahan jumlah set kromosom menyebabkan munculnya individu monoploid (n), triploid (3n), tetraploid (4n), dan seterusnya. Peristiwa euploidi dapat dibedakan menjadi dua, yaitu (1) a**utopoliploidi**, dimana proses poliploidisasinya dilakukan spontan, dan (2) a**llopoliploidi**, yang memerlukan induksi dalam proses poliplodisasinya.

b. Aneuploidi

Aneuploidi merupakan mutasi kromosom yang menyebabkan perubahanjumlah kromosom dalam tubuh individu. Individu normal adalah individu diploid (2n). Manusia normal memiliki 46 kromosom (2n=46). Perubahan jumlah kromosom menyebabkan munculnya individu nullisomi (2n-2), monosomi (2n-1), trisomi (2n+1), tetrasomi (2n+2), dan seterusnya.

2. Mutasi Gen (*Point Mutation*)

Mutasi gen merupakan perubahan urutan sekuens gen karena terjadi perubahan pada level DNA. Mutasi DNA dapat terjadi melalui proses ***transisi, transversi, delesi, insersi,*** dan ***duplikasi***. Mutasi gen dapat berdampak pada terjadinya *silent mutation* (mutasi yang tidak menyebabkan perubahan asam amino), *nonsense mutation* (mutasi yang menyebabkan terbentuknya kodon STOP) dan *missense mutation* (mutasi yang menyebabkan perubahan protein). Mutasi gen dapat dibedakan menjadi dua, yaitu:

a. Pergantian Basa Nitrogen (Substitution Mutation)

Mutasi pergantian basa nitrogen dapat dibedakan menjadi dua, yaitu: (1) *transisi*, apabila basa nitrogen berganti dengan dengan basa nitrogen satu golongan (purin berganti dengan purin, pirimidin berganti dengan pirimidin), dan (2) *transversi*, apabila basa nitrogen berganti dengan dengan basa nitrogen berbeda golongan (purin berganti dengan pirimidin, atau sebaliknya).

b. Perubahan Jumlah Basa Nitrogen (Frameshift Mutation)

Mutasi gen yang berdampak pada perubahan jumlah basa nitrogen dapat terjadi *melalui* proses delesi (pengurangan), insersi (penyisipan), dan duplikasi (penggandaan) DNA.

4.3 Mutasi Pada Sequence

Mutasi adalah perubahan *sequence* genetik, yang merupakan penyebab utama perbedaan di antara organisme. Perubahan ini terjadi pada berbagai tingkatan, dengan konsekuensi yang sangat berbeda. (http:// www.nature.com /scitable/topicpage/genetic-mutation-1127). Menurut Shen (2007), mutasi pada *sequence* DNA dapat diklasifikasikan menjadi 4 tipe, yaitu :

a. Tipe I

   Suatu mutasi yang disebabkan oleh perubahan nukleotida, misalnya dri "a" menjadi "g".

b. Tipe II

   Suatu mutasi yang terjadi karena ada bagian nukleotida yang berubah urutan posisinya, misalnya bagian "accgu" berubah urutan menjadi "guacc".

c. Tipe III

   Suatu mutasi yang disebabkan oleh penyisipan segmen baru ke dalam sequence, misalnya penyisipan "aa" di bagian tengah pada segmen "gguugg" akan mengubah segmen menjadi "gguaaugg"

d. Tipe IV

   Suatu mutasi yang terjadi karena penghapusan segmen nukleotida pada sequence, misalnya menghapus nukleotida "ag" dari segmen "acaguua" sehingga segmen berubah menjadi "acuua".

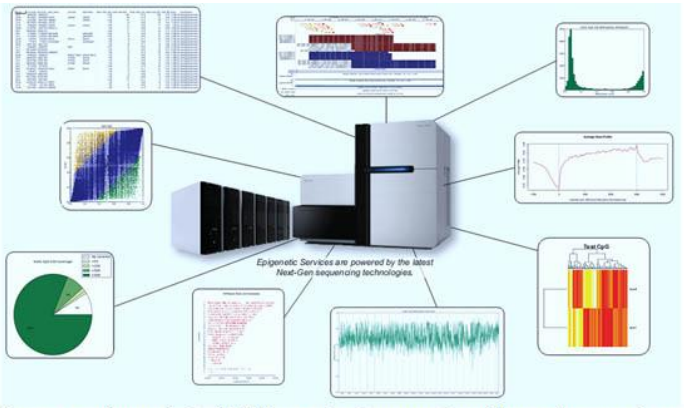
Pada mutasi tipe I dan tipe II, posisi dari semua nukleotida tidak mengalami perubahan maka mutasi ini disebut mutasi subtitusi. Sedangkan untuk mutasi tipe III dan tipe IV yang bisa mengubah posisi nukleotida, maka disebut sebagai mutasi pemindahan.

Bab 5

Penelitian Pesejajaran Sequence di Indonesia

5.1 Bioinformatika sebagai Teknologi Sekuensing Terbaru

Pada masa sekarang ini hingga satu dekade ke depan, para peneliti akan dihadapkan pada tantangan penyimpanan, pengolahan dan analisis data besar (*big data*) yang dihasilkan dari teknologi *Next-Generation Sequencing* (NGS). Analisis sekuen genom menggunakan pendekatan NGS merupakan proses konversi dari materi biologis belum bermakna (sampel DNA atau RNA) menjadi kode informasi (kode biner dan kode basa nukleotida) dan diterjemahkan menjadi informasi biologis bermakna. Bioinformatika terlibat secara nyata dan memiliki peran sangat penting dalam rangka mempermudah setiap tahapan analisis NGS.



Ilustrasi pemanfaatan teknologi NGS dalam analisis ekspresi gen, kuantifikasi metilasi genom hingga pemetaan kromosom. Sumber gambar: http://www.epibeat.com.

Bioinformatika merupakan bagian yang tidak terpisahkan dari teknologi sekuensing generasi baru atau yang sering disebut *Next-Generation Sequencing* (NGS). Pada masa sekarang ini hingga satu dekade ke depan, para peneliti akan dihadapkan pada tantangan penyimpanan, pengolahan dan analisis data besar (*big data*) yang dihasilkan dari teknologi NGS. Setidaknya, empat tahapan wajib dilakukan dalam analisis sekuen nukleotida menggunakan platform NGS yaitu (1) pemanggilan basa (*base calling*) pasca sekuensing, (2) penjajaran dan penggabungan sekuen (*assembly*), (3) anotasi sekuen (*annotation*), dan (4) integrasi data bioinformatika menjadi data biologis (*data integration*).

Pertama, analisis pemanggilan basa (*base calling*) pada potongan pendek sekuen (100-500 pasangan basa) yang disebut *reads* pasca prosesi sekuensing genom atau transkrip dilakukan menggunakan piranti lunak berbasis modul perintah (*Command Line Interface*, CLI). **Kedua**, *contig* dan *scaffold* disusun menggunakan penjajaran dan penggabungan (*assembly*) sekuen nukleotida pendek tersebut. *Contig* dan *scaffold* merupakan bentuk gabungan *reads* yang umumnya memiliki panjang ratusan hingga ratusan ribu pasang basa. **Ketiga**, anotasi dan visualisasi sekuen *contig* dan *scaffold* kemudian dilakukan menggunakan piranti lunak berbasis grafis (*Graphical Processing Unit*, GPU) seperti BLAST2GO. **Keempat**, keseluruhan analisis bioinformatika dari platform NGS seperti GO (Gene Ontology), KEGG (*Kyoto Encyclopedia of Genes and Genomes*) hingga DGE (*Differential Gene Expression*) diintegrasi menjadi satu luaran informasi yang koheren serta memberikan makna biologis. Saat ini, beberapa tool bioinformatika untuk analisis data NGS tersedia secara daring dan tidak dipungut biaya, seperti *Galaxy* (https://usegalaxy.org/) dan *Genomic tools* (http://molbiol-tools.ca/Genomics.htm).

Analisis bioinformatika terlibat bahkan sejak sekuen nukleotida mentah (*raw nucleotide sequence*) dihasilkan dari mesin sekuensing seperti Illumina, SFF, HDF5, CG atau SOLID). Proses sekuensing sendiri mendeteksi basa-basa nukleotida dan mengubahnya secara komputasi menjadi data reads. Sistem format FASTQ digunakan untuk mengukur kualitas dari sekuen reads yang dihasilkan. Pengukuran tersebut pada dasarnya adalah memberikan penilaian apakah basa yang terbaca akurat atau tidak. Data dalam bentuk FASTQ sukar untuk digunakan pada peneliti di laboratorium karena data berukuran besar dan masih berbentuk kode angka dan karakter, bukan dalam bentuk kode basa nukleotida. Oleh sebab itu, data FASTQ pada umumnya dikonversi menjadi bentuk kompak yang disebut SAM (*Sequence Alignment Map*) dan kemudian lebih dikompres menjadi BAM (*Binary Alignment Map*).

Penjajaran sekuen reads untuk membentuknya menjadi sekuen yang lebih panjang berupa contig dan/atau *scaffold* dapat dilakukan dengan dua pendekatan berbasis bioinformatika yaitu (1) pemetaan komparatif *(comparative mapping*) dimana sekuen reads disejajarkan dengan sekuen genom referensi (*reference genome*) dan (2) penggabungan sekuen dengan memanfaatkan sekuen *reads* yang saling tumpang tindih (*overlapping reads*). Pendekatan kedua tersebut lebih dikenal sebagai *de novo assembly*. Saat ini, berbagai genom referensi telah tersedia untuk spesies baik hewan dan tanaman. Umumnya, para peneliti yang melakukan pendekatan *de novo assembly* akan tetap membandingkan dan mengkonfirmasi sekuen mereka menggunakan referensi genom yang telah ada baik pada spesies yang sama maupun pada spesies terdekat. Dalam mengkonfirmasi hasil *assembly* secara manual, para peneliti menggunakan piranti lunak seperti Tablet® untuk memvisualisasikan sekuen *contig* atau *scaffold* mereka.

Pada tahapan berikutnya, bioinformatika dibutuhkan lebih spesifik dalam menganotasi, mengubah dan menerjemahkan sekuen nuleotida menjadi informasi genomika tingkat tinggi seperti menentukan daerah penyandi protein (*coding sequence*, CDS), daerah yang tidak menyandi protein (*non coding*), bentuk isoform sekuen mRNA, sinyal peptida, dan elemen repetitif (*repeat elements*). Pada organisme eukariotik yang memiliki struktur genom lebih kompleks dibanding organisme prokatiotik, anotasi genom menjadi lebih sulit dan menantang. Analisis pada genom yang lebih kompleks umumnya dilakukan menggunakan rangkaian proses anotasi yang disebut GAP (*genome annotation pipeline*). Pada basis data publik, The NCBI Eukaryotic Genome Annotation Pipeline tersedia untuk genom eukariotik (http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/). GAP tersebut terdiri dari langkah pertama berupa identifikasi dan mengeliminasi elemen genom repetitif (mikrosatelit, retrotransposon dan transposon) menggunakan *RepeatMasker*, *Censor* atau *WindowMasker*. Langkah pertama tersebut penting untuk menyaring sekuen repetitif yang dapat mengganggu analisis BLAST dalam anotasi sekuen penyandi protein. Langkah berikutnya dari GAP adalah anotasi transkrip, penjajaran protein/domain, prediksi model gen, penamaan gen dan lokus, dan pemberian GeneID.

Analisis bioinformatika tingkat lanjut pada NGS harus menghasilkan makna biologis dari sekuen genom. Salah satu analisis tingkat lanjut tersebut adalah GO (*Gene Ontology*). GO menyediakan terminologi dari gen, interaksi protein-protein yang terlibat sebagai komponen seluler, fungsi molekuler dan proses biologis terkait. Dalam sejarahnya, analisis GO dimulai hanya dari basis data dari tiga organisme model yaitu *FlyBase* (*Drosophila*), *the Saccharomyces Genome Database* (SGD) *and the Mouse Genome Database* (MGD). Saat ini, Konsorsium Kontributor GO telah terbentuk dan terus menambahkan basis data baru dari organisme yang genomnya baru disekuen. Daftar contributor GO tersebut dapat diakses melalui tautan: http://geneontology.org/page/go-consortium-contributors-list. Selain GO, analisis ontologi serupa juga disediakan oleh beberapa *provider* yang berbeda seperti: *the Open Biological and Biomedical Ontologies (OBBO), Reactome, DAVID, and the KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database*.

Pada akhirnya, analisis sekuen genom menggunakan pendekatan NGS merupakan proses konversi dari materi biologis belum bermakna (sampel DNA atau RNA) menjadi kode informasi (kode biner dan kode basa nukleotida) dan diterjemahkan menjadi informasi biologis bermakna. Bioinformatika terlibat secara nyata dan memiliki peran sangat penting dalam rangka mempermudah setiap tahapan analisis NGS. Bioinformatika terintegrasi dalam NGS telah menjadi bagian tidak terpisahkan dalam rangka memberikan makna biologis pada sekuen.

5.2 Penelitian Analisis Sequence Protein

Pensejajaran Sequence Protein Menggunakan Algoritma Smith Waterman

Dua aspek penting dalam algoritma Smith Waterman ini, antara lain:

a. Menghitung nilai pada tabel dua dimensi
   Algoritma Smith Waterman menmbahkan 0 ketika menghitung $s(i,j)$ sehingga skor negatif tidak akan pernah terjadi pada Algoritma ini. Keuntungannya akan memperjelas lintasan *backward.*

$$s(i,j) = \max \begin{cases} 0 \\ s(i-1,j-1) + s(x_i, y_i) \\ s(i-1,j) - d \\ s(i,j-1) - d \end{cases}$$

b. Algoritma *Traceback*
   Titik awal dan akhir dari metode *backtrace* pada Algoritma Smith Waterman dipilih elemen dengan skor maksimal. Titik akhirnya adalah elemen pertama dengan nilai 0 pada proses *backtrace*. Titik awal dengan skor maksimal akan menjamin skor maksimal pada *alignment sequence* lokal, dan titik akhirnya adalah elemen pertama dengan nilai 0 menjamin bahwa bagian tersebut tidak terlampaui. Bagian yang berhubungan dengan lintasan backward merupakan bagian yang memiliki skor penalti minimum.

Pada pensejajaran ini digunakan 19 sample data sequence protein virus zika, 4 sample data sequence protein pasien terinfeksi virus dengue type 1, type 2, type 3, dan type 4 yang berasal dari Indonesia. Dari keseluruhan sampel virus zika tersebut, diambil 1 sequence virus dari Indonesia (Jambi) dan disejajarkan dengan sequence virus dengue dari masing-masing type, yang juga diambil dari Indonesia. Selanjutnya dicari perbedaan antar sequence, persamaan kedua sequence dan prosentasenya, durasi waktu, dan mutasi yang terjadi. Proses dikerjakan secara statis karena sequence yang disejajarkan hanya 4 sequence dengue dengan 1 sequence zika.

Data sequence virus dengue yang diambil secara online di genbank, database gen terbesar di dunia milik pemerintah Amerika Serikat. Adapun pengambilannya dengan mengakses *National Center for Biotechnology Information* (NCBI, www.ncbi.nlm.nih.gov). Selanjutnya diambil sequence masing-masing virus dan disimpan kode FASTAnya dalam format txt. Adapun data sequence ditunjukkan oleh tabel 1 dan 2 berikut.

Tabel 1 Data sequence protein pasien terinfeksi virus dengue

No	Access Code	Type	Sequence Length	Date of Sample Collection	Explanation
1	AHG06327	1	3392	15-02-2008	Dengue virus 1 isolate Makassar-0398, complete genom
2	AHG06364	2	3391	05-04-2010	Dengue virus 2 isolate Makassar-WS80, complete genom
3	AHG06376	3	3390	22-03-2010	Dengue virus 3 isolate Makassar-WS78, complete genom
4	AHG06382	4	3387	30-04-2008	Dengue virus 4 isolate Makassar-2007, complete genom

Tabel 2 Data sequence protein pasien terinfeksi virus Zika

No	Access Code	Sequence Length	Country	Date of Sample Collection	Explanation
1	KM078936	976	Easter Island Chili	1 Maret 2014	Partial cds
2	KJ873160	893	New Caledonia	3 April 2014	Partial cds
3	KJ776791	10.807	French Polinesia	28 Nov 2013	Complete genom
4	KM851039	789	Thailand	19 Juli 2014	Partial cds
5	KF993678	10.141	Canada	19 Feb 2013	Partial cds
6	AMK 49492	383	Indonesia (Jambi)	30 Des 2014	Partial cds
7	JN860885	10.269	Cambodia	2010	Partial cds
8	EU545988	10.272	Yap Micronesia	Juni 2010	Complete cds
9	KM851038	789	Philippines	9 Mei 2012	Partial cds
10	HQ234499	10.269	Malaysia	1966	Partial cds; host: Aedes Aegypti
11	MR766 /ABY86749	255	EI	2015	Partial cds
12	AY632535 / AAV34151. 1	10.794	Uganda	1947	Complete cds; Host: sentinel monkey
13	KF268948	10.788	Central African Republic	1976	Complete cds; Host: aedes Africanus
14	KF383091	708	Senegal	2001	Partial cds
15	HQ234500	10.251	Nigeria	1968	Partial cds
16	KF383084	708	Senegal	1991	Partial cds
17	HQ234501	10.269	Senegal	1984	Partial cds
18	KF383113	708	Cote de Ivoire	1980	Partial cds
19	DQ859064	10.290	South Africa	-	Complete cds; Spondweni virus

Dengan menggunakan algoritma Smith Waterman, pensejajaran antara virus zika dengan virus Dengue ditunjukkan pada table 3 berikut.

Table 3. Matlab Berdasarkan Algoritma Smith Waterman

Sequence			Similarity/ Dissimilarity		Percentace (%)		Duration (s)
DEN-V	Type	ZIK-V	Sim	Diss	Sim	Diss	
AHG06327	1	AMK49492	273	108	71.4660	28.27	0.062
AHG06364	2	AMK49492	272	108	71.0183	28.20	0.094
AHG06376	3	AMK49492	274	107	71.5405	27.94	0.359
AHG06382	4	AMK49492	271	110	70.7572	28.72	0.156

Berdasarkan keluaran empat pensejajaran, dan dibandingkan dengan Basic Search Alignment Search Tool (BLAST), yaitu sebuah program untuk membandingkan urutan nukleotida atau protein ke database urutan dan menghitung signifikansi statistik dari dua urutan kecocokan ditunjukkan pada tabel 4 berikut.

Tabel 4. Perbandingan Matlab and BLAST

Sequence			Sequence length		Identical value		Duration (s)	
DEN-V	Type	ZIK-V	DEN-V	ZIK-V	Matlab	BLAST	Matlab	BLAST
AHG06327	1	AMK49492	3392	383	71,466 %	71 %	0,062	12,16
AHG06364	2	AMK49492	3391	383	71,0183%	71 %	0,094	11,27
AHG06376	3	AMK49492	3390	383	71,5405%	71 %	0,359	6,58
AHG06382	4	AMK49492	3387	383	70,7572%	71 %	0,156	10,68

Dari tabel 4 di atas, dapat diketahui bahwa output pensejajaran sequence virus Zika dan virus Dengue pada simulasi matlab memiliki nilai lebih lebih akurat daripada output BLAST. Terbukti dengan tingkat keakuratan output simulasi matlab sampai 4 angka desimal, sedangkan BLAST hanya menampilkan 2 angka signifikan dan juga durasi waktu komputasi pada simulasi matlab lebih pendek dari waktu komputasi pada BLAST (Pradana & Amiroch, 2018).

5. Pohon Filogenetik Penyebaran Virus Zika Jambi

Analisis Penyebaran Virus Zika berdasarkan output MUSCLE

Sebelum dikonstruksi pohon filogenetik, masing-masing sequence disejajarkan terlebih dahulu. Dari 19 data yang diperoleh, ternyata jenis datanya tidak sama, beberapa menggunakan data protein, beberapa data DNA. Sehingga untuk keseragaman, sebagian besar data DNA diubah menjadi data protein dan diakses mengikuti penamaan sequence proteinnya. Sehingga data protein keseluruhan sequence virus Zika tersebut ditunjukkan tabel 5 berikut.

Tabel 5. Kode akses protein yang digunakan sebagai data

No	Kode semula	Kode akses Protein	Panjang seq. (bp)	Negara	Tanggal pengumpulan sampel
1	KM078936	AJD79008	976	Easter Island Chili	1 Maret 2014
2	KJ873160	AJA40023	893	New Caledonia	3 April 2014
3	KJ776791	AHZ13508	10.807	French Polinesia	28 Nov 2013
4	KM851039	AKH87423	789	Thailand	19 Juli 2014
5	KF993678	AHL37808	10.141	Canada	19 Feb 2013
6	AMK 49492	AMK49492	383	Indonesia (Jambi)	30 Des 2014
7	JN860885	AFD30972	10.269	Cambodia	2010
8	EU545988	ACD75819	10.272	Yap Micronesia	Juni 2010
9	KM851038	AKH87424	789	Philippines	9 Mei 2012
10	HQ234499	AEN75264	10.269	Malaysia	1966
11	MR766 /ABY86749	ABY86749	255	EI	2015
12	AY632535 / AAV34151.1	AAV34151	10.794	Uganda	1947
13	KF268948	AHF43978	10.788	Central African Republic	1976
14	KF383091	AHL43476	708	Senegal	2001
15	HQ234500	AEN75265	10.251	Nigeria	1968
16	KF383084	AHL43469	708	Senegal	1991
17	HQ234501	AEN75266	10.269	Senegal	1984

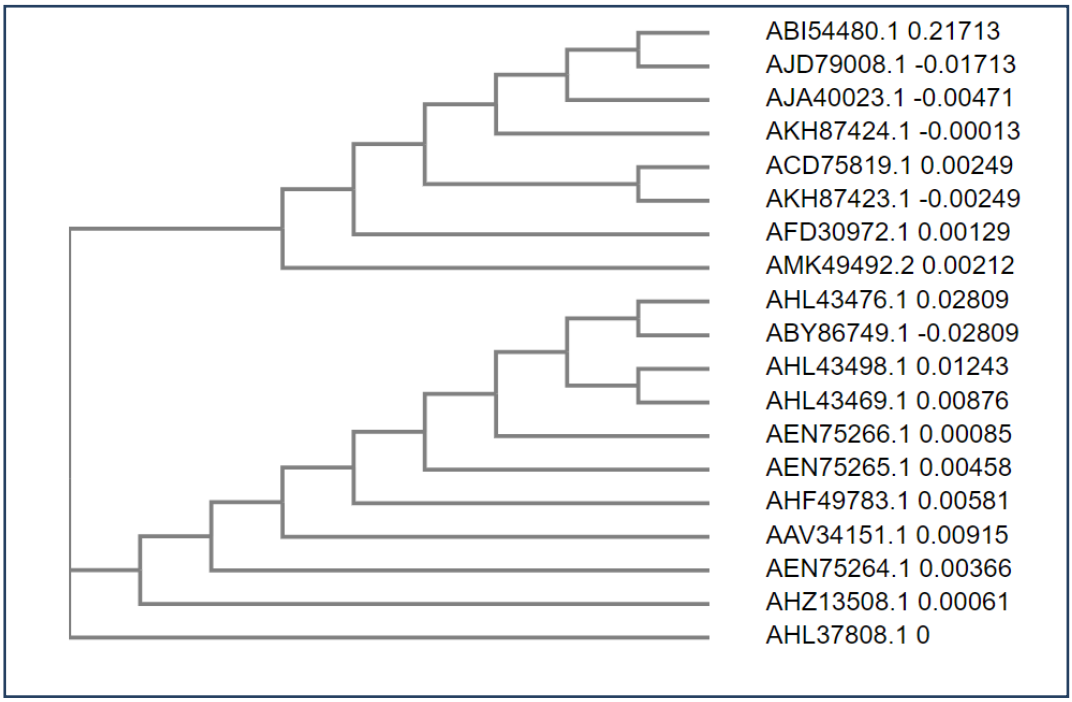
18	KF383113	AHL43498	708	Cote de Ivoire	1980
19	DQ859064	ABI54480	10.290	South Africa	-

Dari output CLUSTALW2 diperoleh matriks prosentase identik pada masing-masing sequence seperti berikut ini.

Matrix nilai identik dari 14 sequence

1: ABI54480.1	100.00	76.69	74.97	77.95	71.76	75.00	75.04	75.03	74.97	74.85	77.57	80.00	78.11	76.69	74.91	77.97	75.20	75.04	75.12
2: AHL43476.1	76.69	100.00	91.95	90.40	-nan	90.68	90.68	90.68	90.68	90.68	89.83	90.68	90.71	93.64	93.64	94.92	93.62	94.07	94.49
3: AEN75264.1	74.97	91.95	100.00	98.48	96.86	98.92	98.91	98.69	98.86	98.57	98.10	97.85	97.64	94.07	97.08	94.49	97.60	97.16	97.40
4: AKH87424.1	77.95	90.40	98.48	100.00	-nan	99.62	99.62	99.62	99.62	99.62	99.62	99.61	99.60	94.92	96.20	94.35	96.18	96.58	96.58
5: ABY86749.1	71.76	-nan	96.86	-nan	100.00	96.47	96.08	96.47	96.47	96.47	-nan	100.00	-nan	-nan	97.65	-nan	98.43	98.43	98.43
6: AHZ13508.1	75.00	90.68	98.92	99.62	96.47	100.00	99.88	99.65	99.65	99.18	100.00	100.00	100.00	92.80	96.61	93.22	97.25	96.81	97.11
7: AHL37808.1	75.04	90.68	98.91	99.62	96.08	99.88	100.00	99.70	99.70	99.41	100.00	100.00	100.00	92.80	96.80	93.22	97.31	96.83	97.16
8: AMK49492.2	75.03	90.68	98.69	99.62	96.47	99.65	99.70	100.00	99.47	99.01	100.00	100.00	100.00	92.80	96.40	93.22	97.02	96.58	96.87
9: AFD30972.1	74.97	90.68	98.86	99.62	96.47	99.65	99.70	99.47	100.00	99.24	100.00	100.00	100.00	92.80	96.46	93.22	97.11	96.66	96.96
10: ACD75819.1	74.85	90.68	98.57	99.62	96.47	99.18	99.41	99.01	99.24	100.00	100.00	100.00	100.00	92.80	96.49	93.22	96.81	96.34	96.67
11: AKH87423.1	77.57	89.83	98.10	99.62	-nan	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	94.35	95.82	93.79	95.80	96.20	96.20
12: AJD79008.1	80.00	90.68	97.85	99.61	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	92.80	95.38	93.22	95.37	95.69	95.69
13: AJA40023.1	78.11	90.71	97.64	99.60	-nan	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	92.92	94.95	93.36	94.93	95.29	95.29
14: AHL43498.1	76.69	93.64	94.07	94.92	-nan	92.80	92.80	92.80	92.80	92.80	94.35	92.80	92.92	100.00	97.46	97.88	97.02	97.88	98.31
15: AAV34151.1	74.91	93.64	97.08	96.20	97.65	96.61	96.80	96.40	96.46	96.49	95.82	95.38	94.95	97.46	100.00	97.88	98.51	98.07	98.48
16: AHL43469.1	77.97	94.92	94.49	94.35	-nan	93.22	93.22	93.22	93.22	93.22	93.79	93.22	93.36	97.88	97.88	100.00	97.45	98.31	98.73
17: AHF49783.1	75.20	93.62	97.60	96.18	98.43	97.25	97.31	97.02	97.11	96.81	95.80	95.37	94.93	97.02	98.51	97.45	100.00	98.57	98.98
18: AEN75265.1	75.04	94.07	97.16	96.58	98.43	96.81	96.83	96.58	96.66	96.34	96.20	95.69	95.29	97.88	98.07	98.31	98.57	100.00	99.03
19: AEN75266.1	75.12	94.49	97.40	96.58	98.43	97.11	97.16	96.87	96.96	96.67	96.20	95.69	95.29	98.31	98.48	98.73	98.98	99.03	100.00

Terlihat bahwa nilai identiknya tinggi sekali. Hal itu menunjukkan bahwa tingkat kemiripan sequence satu dengan yang lain sangatlah tinggi. Beberapa sequence menunjukkan kemiripan 100% seperti pada sequence 5 dan sequence 12, sequence 6 dengan sequence 11,12,13, dan seterusnya (Pradana & Amiroch, 2018). Untuk pohon filogenetik bentukan MUSCLE tampak pada gambar 1 berikut.



Gambar 1. Pohon filogenetik hasil output program MUSCLE

Berdasarkan gambar 4.5 tersebut, diketahui bahwa virus Zika terpisah menjadi dua cluster. Untuk lebih jelasnya dibentuk tabel 6 berikut.

Tabel 6. Cluster I penyebaran Virus Zika

No.	Kode akses	Nama daerah
1	ABI54480	South Africa
2	AJD79008	Easter Island Chili
3	AJA40023	New Caledonia
4	AKH87424	Philipina
5	ACD75819	Yap Micronesia
6	AKH87423	Thailand
7	AFD30972	Cambodia
8	AMK49492	Indonesia (Jambi)

Karena sebagian besar penyebaran virus pada cluster I ini berada di Asia, maka disebut "Cluster Asia". Nampak pada pohon filogenetik penyebaran hingga ke Indonesia yang pertama kali berasal dari Afrika Selatan, menyebar ke pulau Chili, Caledonia, Philipina, Yap Micronesia, Thailand, Cambodia, dan akhirnya sampai ke Indonesia (Pradana & Amiroch, 2018). Sedangkan untuk cluster kedua, daerah penyebarannya ditabelkan sebagai berikut:

Tabel 7. Cluster II penyebaran Virus Zika

No.	Kode akses	Nama daerah
1	AHL43476	Senegal
2	ABY86749	El-Salvador
3	AHL43498	Cote de Ivoire
4	AHL43469	Senegal
5	AEN75266	Senegal
6	AEN75265	Nigeria
7	AHF49783	Central African Republic
8	AAV34151	Uganda
9	AEN75264	Malaysia
10	AHZ13508	French Polinesia

Dari cluster II pada tabel 7 tersebut, sebagian besar berada pada daerah Afrika, sehingga cluster kedua disebut “Cluster Africa”.

DAFTAR PUSTAKA

Aprijani, D.A., & Elfaizi, MA. (2004). Bioinformatika Perkembangan, Disiplin Ilmu dan Penerapannya di Indonesia.

Horner DS, et al.(2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Briefings in Bioinformatics. 11(2):181-97.

Milne I. (2016). Tablet: Visualizing Next-Generation Sequence Assemblies and Mappings. In: Edwards D, editor. Plant Bioinformatics: Methods and Protocols. New York, NY: Springer New York;. p. 253-68.

Pradana, M. S., & Amiroch, S. (2018). Zika Virus Mutation and The Spreading to Indonesia. *International Journal of Computing Science and Applied Mathematics*, *4*(1), 15–18.

www.ncbi.nlm.nih.gov