**TUGAS RESUME**

**DATA MINING**

Tugas ini disusun untuk Memenuhi tugas mata kuliah metodologi pembelajaran

**Dosen pengampu: Dr. Wahyudi Setiawan, M. Pd.**

****

**Disusun Oleh**

**Giraldo Nainggolan (220441100064)**

**Wisnu Ary Swadana (220441100121)**

**Abib Maulana Aan Nafudi (220441100118)**

**Fairuz Abdullah (220441100070)**

**Program Studi Sistem Informasi**

**Fakultas Teknik**

**Universitas Truojoyo Madura**

**Tahun Ajaran 2024 / 2025**

**CDD/SPARCLE: the conserved domain**

**database in 2020**

**Sumber**: Ju Nucleic Acids Research, 2020, Vol. 48, Database https://academic.oup.com/nar/article/48/D1/D265/5645006#190998879

**Penulis**: issue Shennan Lu, Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, David I. Hurwitz, Gabriele H. Marchler, James S. Song, Narmada Thanki, Roxanne A. Yamashita, Mingzhang Yang, Dachuan Zhang, Chanjuan Zheng, Christopher J. Lanczycki and Aron Marchler-Bauer

**Abstrak**

Saat Conserved Domain Database (CDD) NLM memasuki tahun ke-20 operasinya sebagai sumber daya yang tersedia untuk umum, staf kurasi CDD terus mengembangkan klasifikasi hirarkis keluarga domain protein yang tersebar luas, dan mencatat situs yang dilestarikan yang terkait dengan fungsi molekuler, sehingga dapat dipetakan ke dalam kueri pengguna untuk mendukung penelitian biomolekuler yang digerakkan oleh hipotesis. CDD menawarkan arsip anotasi domain yang telah dikomputasi sebelumnya serta layanan pencarian langsung untuk kueri protein atau nukleotida tunggal dan kumpulan sekuens kueri protein yang lebih besar. Staf CDD terus mengkarakterisasi keluarga protein melalui arsitektur domain yang dilestarikan dan telah membangun korpus yang signifikan dari arsitektur domain yang dikuratori untuk mendukung penamaan protein bakteri di RefSeq. Definisi arsitektur ini tersedia melalui SPARCLE, Mesin Pelabelan Arsitektur Protein Subfamili. CDD dapat diakses di https://www.ncbi.nlm.nih.gov/Structure/ cdd/cdd.shtml.

**Pendahuluan**

Klasifikasi data adalah proses pengkategorian data ke dalam kelompok-kelompok yang bermakna. Klasifikasi data banyak digunakan di berbagai bidang, seperti pendidikan, asuransi, media sosial, dan pemasaran.

Pada saat penulisan, CDD versi v3.17 adalah versi produksi langsung dengan 52.910 model protein dan domain protein yang diperoleh dari Pfam (1), SMART (2), koleksi COG (3), TIGRFAMS (4), koleksi NCBI Protein Cluster (5), NCBIfam (6), dan upaya kurasi data internal CDD (7). Versi CDD v3.18 akan dirilis pada musim dingin 2019/2020 dan akan menyertakan Pfam versi 32 dan total 55.434 model protein dan domain protein. Untuk CDD v3.18, ukuran

ukuran koleksi model saat ini, sehingga menghasilkan nilai E yang sedikit lebih tinggi yang dilaporkan oleh RPS-BLAST (8). Koleksi NCBIfam dalam CDD adalah sekumpulan model yang berasal dari HMM yang telah dikembangkan untuk meningkatkan anotasi genom bakteri. Saat ini, CDD tidak menyertakan model NCBIfam yang dibuat untuk mengidentifikasi protein yang terlibat dalam resistensi antimikroba, karena cakupannya yang sempit. Tabel 1 menunjukkan 20 klasifikasi terbesar untuk keluarga domain yang umum dan beragam secara fungsional yang baru-baru ini diperbarui atau ditambahkan ke CDD. Secara keseluruhan, lebih dari 4.700 model yang dikurasi oleh grup CDD telah diterbitkan atau diperbarui sejak rilis CDD v3.16. Pada saat rilis CDD v3.17, CDD telah menganotasi sekitar 85% sekuens dalam basis data Entrez/protein (tidak termasuk sekuens dari pengambilan sampel lingkungan). CDD juga mencakup sekitar 94% sekuens protein (lebih panjang dari 30 residu) yang berasal dari struktur 3 dimensi protein yang disediakan melalui MMDB. Staf kurasi CDD memantau sekuens turunan struktur yang belum tercakup dalam CDD untuk keluarga domain protein baru dengan distribusi taksonomi yang luas dan menghasilkan model keluarga domain yang sesuai secara de novo. Untuk CDD v3.18, total 33.980 anotasi situs tersedia pada 12.418 dari 16.069 model domain yang dikurasi oleh staf CDD. Pola urutan telah dicatat untuk 3250 anotasi situs ini, sehingga kecocokan pola menentukan apakah sebuah

**Kajian Teori**

**Klasifikasi Data**

Klasifikasi data adalah proses pengkategorian data ke dalam kelompok-kelompok yang bermakna. Ada banyak metode klasifikasi data, seperti machine learning, deep learning, dan statistik.

**Pembelajaran mesin**

Machine learning adalah algoritma yang dapat belajar dari data dan meningkatkan kinerjanya tanpa diprogram secara eksplisit.expand\_more Machine learning dapat digunakan untuk klasifikasi data dengan menganalisis data dan menemukan pola yang dapat digunakan untuk memprediksi kategori data baru.expand\_more

**Pembelajaran Mendalam**

Deep learning adalah subset dari machine learning yang menggunakan jaringan saraf tiruan untuk belajar dari data.expand\_more Deep learning dapat digunakan untuk klasifikasi data dengan

menganalisis data dan menemukan pola yang kompleks yang tidak dapat ditemukan dengan metode machine learning tradisional.expand\_more.

**Statistik**

Statistik dapat digunakan untuk klasifikasi data dengan menganalisis data dan menemukan distribusi data. Distribusi data dapat digunakan untuk memprediksi kategori data baru.

**Pembahasan**

Layanan TURR (Ture Retrieval Tool) (9) mengelompokkan protein dalam basis data Entrez berdasarkan arsitektur superfamili domain umum. Di sisi lain, SPARCLE untuk 'Mesin Pelabelan Arsitektur Protein Subfamili' mengelompokkan protein berdasarkan SDA, dan kami telah terlibat dalam upaya kurasi yang mengulas SDA yang terwakili dengan baik dalam koleksi sekuens protein dan mengaitkannya dengan saran nama protein dan deskripsi fungsional singkat. Hingga saat ini, kurator CDD telah memberikan nama dan label fungsional pada ∼25.000 SDA,

dengan fokus pada SDA yang umum dalam genom bakteri. Basis data Entrez yang dapat diakses secara publik mendukung kueri teks dan menunjukkan ringkasan informasi untuk SDA serta tautan ke basis data lain, yang terpenting adalah koleksi protein NCBI. CD-Search tidak hanya menampilkan anotasi domain dan fitur, tetapi juga nama dan fungsi.

**Simpulan**

Data tersebut mencantumkan URL untuk layanan, alat bantu, dan koleksi data yang disediakan oleh CDD. RPS-BLAST adalah bagian dari distribusi perangkat lunak BLAST NCBI. Basis data pencarian RPS-BLAST yang telah diformat sebelumnya tersedia sehingga pencarian domain yang dikonservasi dapat dijalankan secara lokal, dan hasilnya dapat diformat dengan utilitas rpsbproc sehingga sesuai dengan laporan yang dihasilkan oleh CD-Search (10) dan BATCH CD-Search, termasuk anotasi situs. Sebuah utilitas baru, sparclbl (SparcleLabel) tersedia melalui FTP; sparclbl memproses hasil dari pencarian RPS-BLAST lokal dan memberikan saran untuk nama protein berdasarkan arsitektur domain. Versi internal sparclbl merupakan bagian dari pipa anotasi genom prokariotik NCBI (PGAP) (6).

**Saran**

* Tampilan flatfile (Format GenPept) dari protein bakteri harusnya di ambil dari data koleksi RefSeq karna lebih akurat dan sama pengelompokannya.
* Untuk awalan aksesksi Bakteri pada protein diubah menjadu "WP" karna memudahkan pembeda jenisnya hannya (disorot dengan oval merah).
* Model Markov Hidden dan aturan BLAST harusnya datanya jangan tersebunyi supaya bisa bermanfaat dicari data basenya untuk publik