

Protein information resource

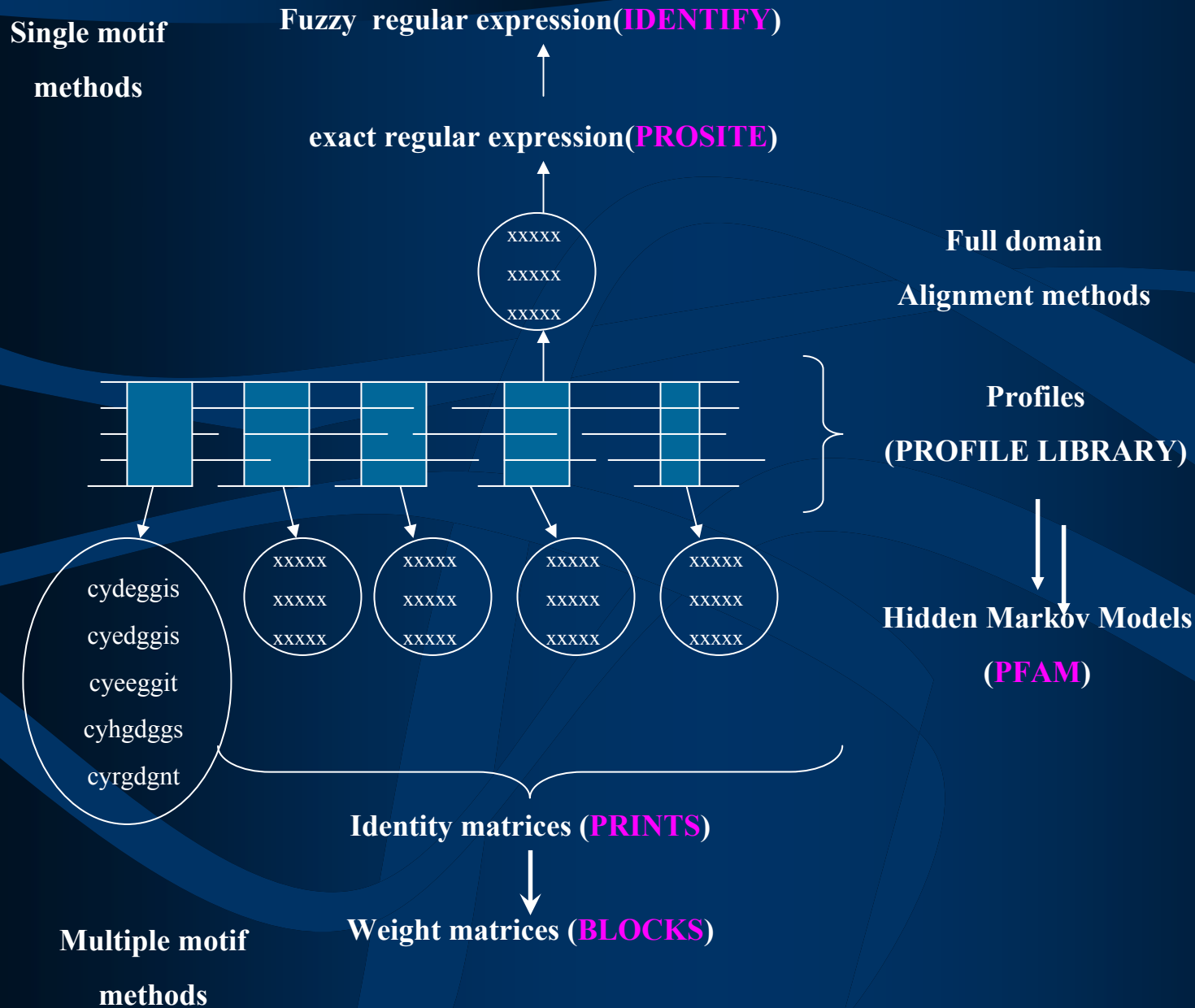
Summary

Secondary Database

Secondary database	Primary source	Stored information
PROSITE	SWISS-PROT	Regular expressions(patterns)
Profiles	SWISS-PROT	Weighted matrices(profiles)
<u>PRINTS</u>	OWL*	<u>Aligned motifs</u> (fingerprints)
Pfam	SWISS-PROT	Hidden Markov Models(HMMs)
BLOCKS	PROSITE/PRINTS	Aligned motifs(blocks)
IDENTIFY	BLOCKS/PRINTS	Fuzzy regular expressions(patterns)

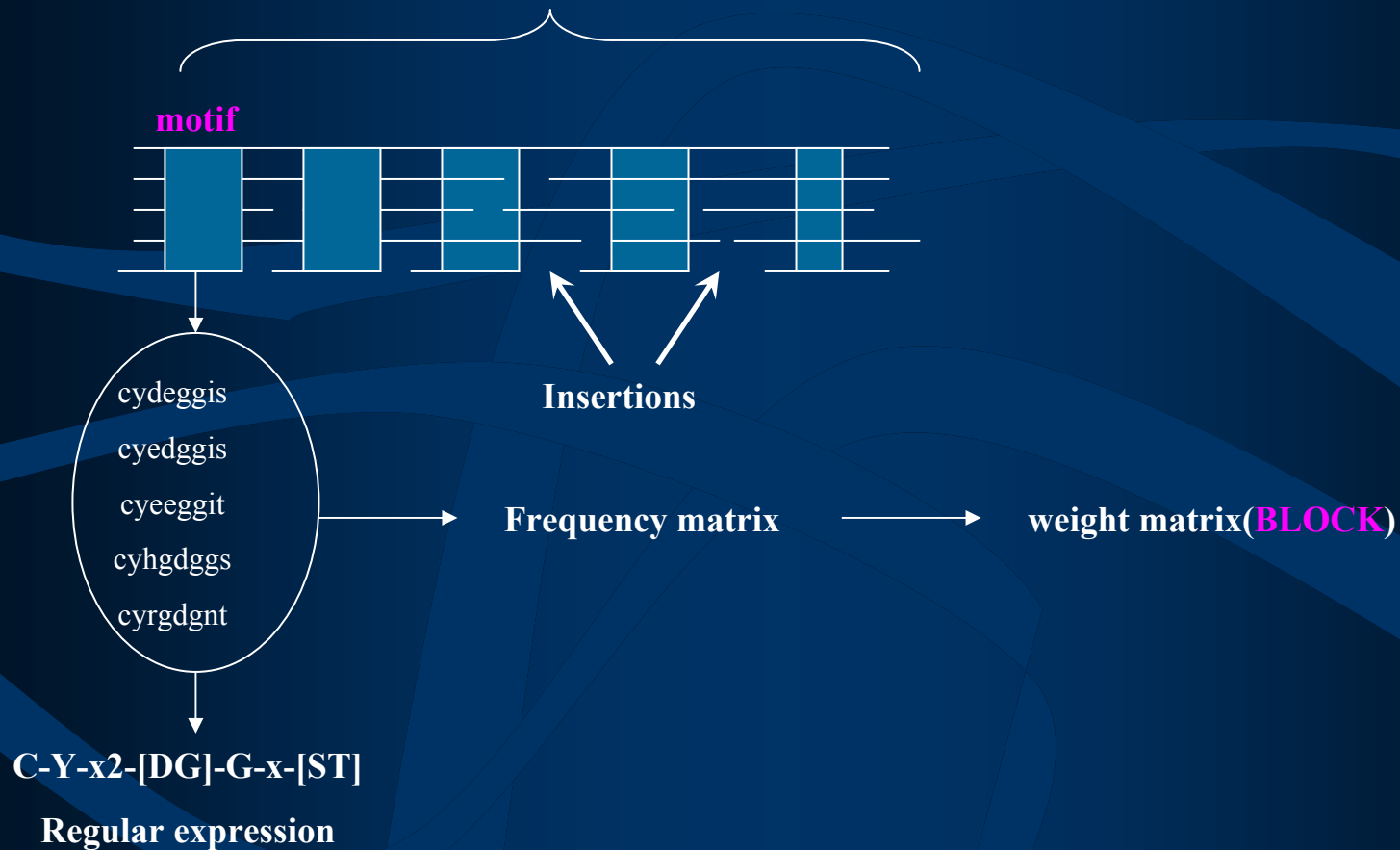
*SWISS-PORT is OWL's highest priority source.

Principal methods for building pattern Database



Terms used in sequence analysis methods

Identity matrices (**PRINTS**)



- Databases are used to store the vast amounts of information issuing from the genome projects.
- There are many different types of database, but for routine protein sequence analysis, **primary, secondary and composite databases are initially the most important.**
- Primary databases contain sequence data(**nucleic acid or protein**) **SWISS-PROT** is the most highly annotated protein sequence database.

- Composite databases amalgamate a variety of different primary sources and are hence efficient to search because they obviate the need to interrogate multiple resources.
- Different composite databases use different primary sources and different redundancy criteria in their amalgamation procedures.

- **Secondary databases** contain **pattern data**, i.e., diagnostic **signatures for protein families**. These signatures encode the most highly conserved features of multiply aligned sequences, which are often **crucial to the structure or function of the protein**.
- Different sequence analysis methods have given rise to different pattern database: the main approaches exploit **single motifs** (e.g., **regular expressions**), **multiple motifs** (e.g., **fingerprints**) and **full domain alignments** (e.g., **Hidden Markov Models**).

- There is some **overlap** in content between the secondary databases, but even their combined total is well short of the estimated **1000~10000 protein families**. Pattern database growth is **slow** because the addition of detailed family annotation is very **time consuming**. **PROSITE** and **PRINTS** are the only comprehensively, manually annotated secondary databases.

- To address the annotation **bottleneck**, the secondary database curators are together **creating a unified database** of protein families known as **InterPro**. It is hoped that **InterPro** will make sequence analysis more straightforward in the future.

- 國內核酸&蛋白質整合型軟體：
GCG (Genetics Computer Group)
<http://gcg.nhri.org.tw/>
- SeqWeb 為 GCG 為改善使用者介面所發展而來的軟體，使用者可直接在瀏覽器上以點選的方式進行序列的比對及分析。
- GenWeb 快速序列搜尋系統，是專為序列搜尋所設計的，在執行序列搜尋時的速度可以比一般系統快上數十至數百倍。

- Protein **structure-related resources** may provide a useful adjunct to the database available for sequence analysis, and will become **increasingly important as more 3D structures are solved.**