HIDDEN MARKOV MODELS AND THE LANGUAGE OF THE PROTEIN UNIVERSE

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

CB_23_24_L33&34 , Rome 18 and 19 Dec 2023

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

Introduction to Hidden Markov Models

From a set of slides by Pietro Liò

(notation same as in Higgs and Attwood)

- About letters, alphabets and states
- First order Markov models
- Higher order models
- Hidden Markov models
- Evaluation Problem
- Decoding Problem
- Learning problem

- Used in computational structural biology
- e.g. to associate to each family of functionally similar proteins
- a generative probabilistic model as a fingerprint

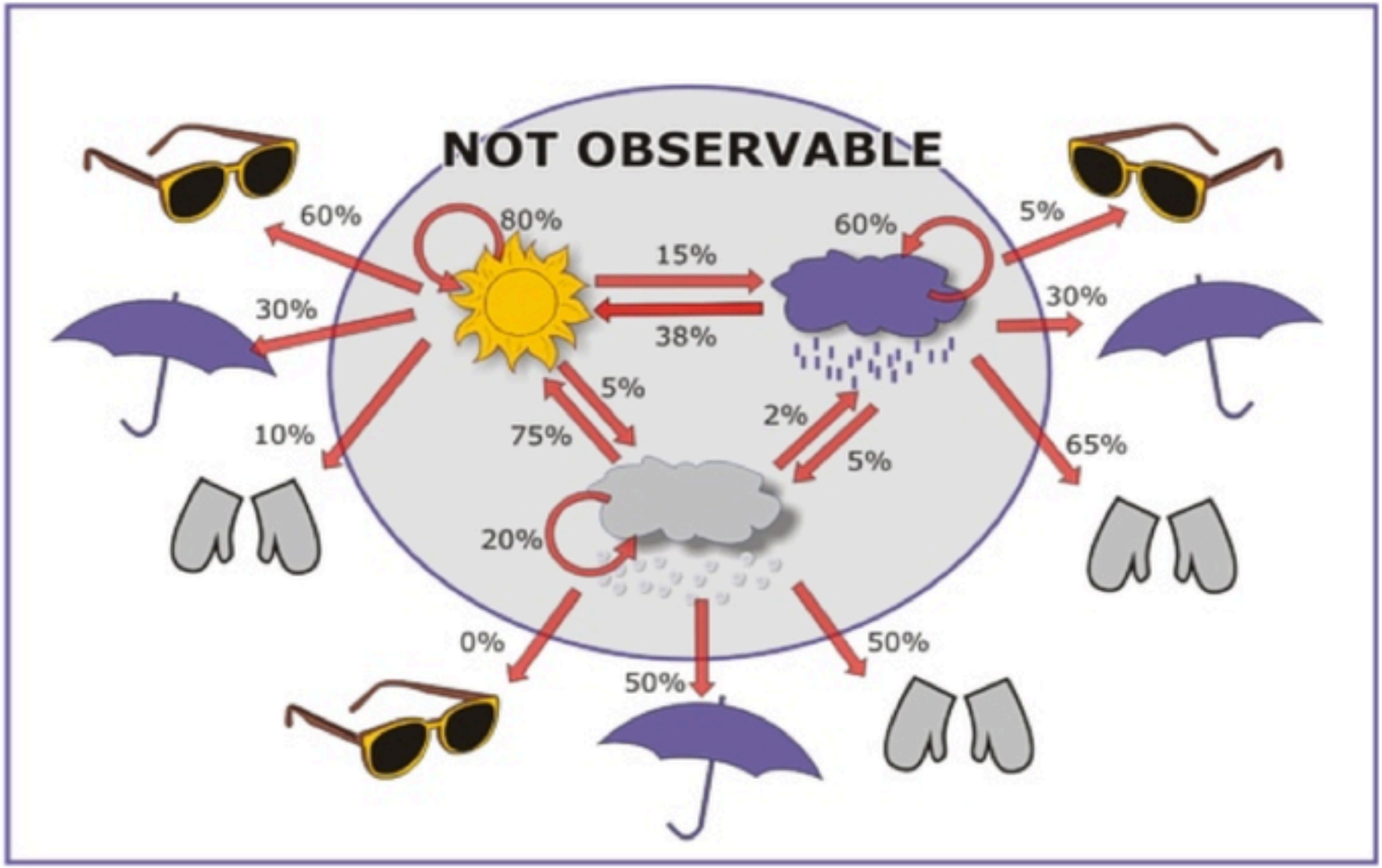
Fasta Format

```
>gi|18089116|gb|BC020718.1| Homo sapiens I factor
AAATTTCAAAAGAATACCTGGAGTGGAAAAGAGTTCTCAGCAGAGACAAAGACCCCGAACACCTCCAACA
TGAAGCTTCTTCATGTTTTCCTGTTATTTCTGTGCTTCCACTTAAGGTTTTGCAAGGTCACTTATACATC
TCAAGAGGATCTGGTGGAGAAAAAGTGCTTAGCAAAAAAATATACTCACCTCTCCTGCGATAAAGTCTTC
TGCCAGCCATGGCAGAGATGCATTGAGGGCACCTGTGTTTGTAAACTACCGTATCAGTGCCCAAAGAATG
GCACTGCAGTGTGTGCAACTAACAGGAGAAGCTTCCCAACATACTGTCAACAAAAGAGTTTGGAATGTCT
TCATCCAGGGACAAAGTTTTTAAATAACGGAACATGCACAGCCGAAGGAAAGTTTAGTGTTTCCTTGAAG
CATGGAAATACAGATTCAGAGGGAATAGTTGAAGTAAAACTTGTGGACCAAGATAAGACAATGTTCATAT
GCAAAAGCAGCTGGAGCATGAGGGAAGCCAACGTGGCCTGCCTTGACCTTGGGTTTCAACAAGGTGCTGA
TACTCAAAGAAGGTTTAAGTTGTCTGATCTCTCTATAAATTCCACTGAATGTCTACATGTGCATTGCCGA
GGATTAGAGACCAGTTTGGCTGAATGTACTTTTACTAAGAGAAGAACTATGGGTTACCAGGATTTCGCTG
ATGTGGTTTGTTATACACAGAAAGCAGATTCTCCAATGGATGACTTCTTTCAGTGTGTGAATGGGAAATA
CATTTCTCAGATGAAAGCCTGTGATGGTATCAATGATTGTGGAGACCAAAGTGATGAACTGTGTTGTAAA
GCATGCCAAGGCAAAGGCTTCCATTGCAAATCGGGTGTTTGCATTCCAAGCCAGTATCAATGCAATGGTG
AGGTGGACTGCATTACAGGGGAAGATGAAGTTGGCTGTGCAGGCTTTGCATCTGTGGCTCAAGAAGAAAC
AGAAATTTTGACTGCTGACATGGATGCAGAAAGAAGACGGATAAAATCATTATTACCTAAACTATCTTGT
GGAGTTAAAAACAGAATGCACATTCGAAGGAAACGAATTGTGGGAGGAAAGCGAGCACAACTGGGAAAAA
TGAAGCAAATCTCATTGGATATTTTTAAAGGTCTCCACAGAGTTTATGCCATATTGGAATTTTGTTGTAT
AATTCTCAAATAAATATTTTGGTGAAGCCAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

Probability of a Sequence of Events

You go out on a certain day and you (...crazy?) and start to keep a record of the habits of people you come cross with: do they wear sunglasses? Do they wear gloves? Do they brandish an umbrella? You get a sequence of events...

Hidden Markov Models



Refresh: definition of a HMM

Definition: A hidden Markov model (HMM)

- Alphabet $\Sigma = \{ b_1, b_2, \ldots, b_M \}$
- Set of states $Q = \{ 1, \ldots, K \}$
- Transition probabilities between any two states

  a_{ij} = transition prob from state i to state j

  $a_{i1} + \ldots + a_{iK} = 1$, for all states i = 1…K

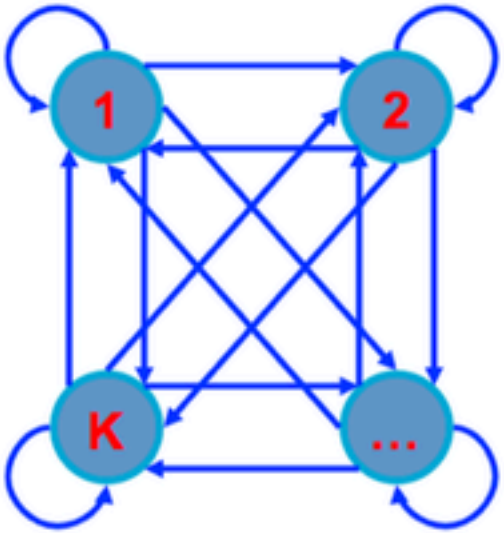
- Start probabilities a_{0i}

  $a_{01} + \ldots + a_{0K} = 1$

- Emission probabilities within each state

  $e_i(b) = P(x_i = b \mid \pi_i = k)$

  $e_i(b_1) + \ldots + e_i(b_M) = 1$, for all states i = 1…K



The three main questions on HMMs

1. Evaluation

   GIVEN a HMM M, and a sequence x,

   FIND Prob[x | M]

2. Decoding

   GIVEN a HMM M, and a sequence x,

   FIND the sequence π of states that maximizes P[x, π | M]

3. Learning

   GIVEN a HMM M, with unspecified transition/emission probs.,
   and a sequence x,

   FIND parameters $\theta = (e_i(.), a_{ij})$ that maximize P[x | θ]

- Hidden Markov Models from proteins to GW

Hidden Markov model tracking of continuous gravitational waves from a neutron star with wandering spin. III. Rotational phase tracking

A. Melatos,[1,2,*] P. Clearwater,[1,2,3] S. Suvorova,[1,2,4,5] L. Sun,[1,2,6,7] W. Moran,[4,5] and R. J. Evans[2,4]

[1] *School of Physics, University of Melbourne, Parkville, Victoria 3010, Australia*
[2] *Australian Research Council Centre of Excellence for Gravitational Wave Discovery (OzGrav), University of Melbourne, Parkville, Victoria 3010, Australia*
[3] *Data61, Commonwealth Scientific and Industrial Research Organisation, Corner Vimiera & Pembroke Roads, Marsfield, NSW 2122, Australia*
[4] *Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3010, Australia*
[5] *School of Electrical and Computer Engineering, RMIT University, Melbourne, Victoria 3000, Australia*
[6] *LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*
[7] *OzGrav-ANU, Centre for Gravitational Astrophysics, College of Science, Australian National University, Australian Capital Territory 2601, Australia*

(Dated: July 28, 2021)

A HMM is a probabilistic finite state automaton defined by a hidden (unobservable) state variable, $q(t)$, and an observable state variable, $o(t)$. The automaton jumps through a time-ordered sequence of observations, $O = \{o(t_0), \ldots, o(t_{N_T})\}$, at discrete times $t_0 \leq \cdots \leq t_{N_T}$. In general there exist $N_Q^{N_T+1}$ possible hidden-state paths, $Q = \{q(t_0), \ldots, q(t_{N_T})\}$, which are consistent with O. Here N_Q counts the finite number of discrete values, that $q(t)$ can take at time t.

Given O, some paths are more likely than others. If we assume that the automaton is Markovian, such that the transition probability from $q(t_n)$ to $q(t_{n+1})$ depends only on $q(t_n)$, then the probability that Q gives rise to O equals

$$\begin{aligned}\Pr(Q|O) = & L_{o(t_{N_T})q(t_{N_T})} A_{q(t_{N_T})q(t_{N_T-1})} \times \ldots \\ & \times L_{o(t_1)q(t_1)} A_{q(t_1)q(t_0)} \Pi_{q(t_0)} \, . \end{aligned} \tag{1}$$

In (1),

$$A_{q_j q_i} = \Pr[q(t_{n+1}) = q_j | q(t_n) = q_i] \tag{2}$$

is the transition probability matrix;

$$L_{o_j q_i} = \Pr[o(t_n) = o_j | q(t_n) = q_i] \tag{3}$$

is the emission probability matrix, namely the probability that the system is observed in state $o(t_n)$ while occupying the hidden state $q(t_n)$; and

$$\Pi_{q_i} = \Pr[q(t_0) = q_i] \tag{4}$$

is the prior vector, namely the probability that the system occupies the hidden state $q(t_0)$ initially.

To solve the HMM, one seeks the most probable path $Q^*(O)$, which maximizes $\Pr(Q|O)$ given O, viz.

$$Q^*(O) = \arg\max \Pr(Q|O) \ . \tag{5}$$

The maximization can be done in many ways. In previous gravitational wave applications as well as in this paper, we employ the Viterbi algorithm, [14, 15] whose logic and pseudocode are summarized briefly in Appendix A. The Viterbi algorithm is a dynamic programming algorithm. It is computationally efficient, executing of order $(N_T + 1)N_Q \ln N_Q$ floating point operations.

Bellman's principle of optimality

Appendix A: Viterbi algorithm

The Viterbi algorithm prunes the tree of possible hidden state sequences Q by appealing to Bellman's Principle of Optimality: if a subpath $\{q^*(t_i), \cdots, q^*(t_j)\}$ is optimal, then all of its subpaths are optimal as well. [80] Dynamic programming is exploited to implement the Principle of Optimality in an efficient, recursive fashion. [14, 15, 18] Pseudocode describing the implementation is presented below in abridged form for ease of reference.

Ingredients of the Viterbi algorithm

At time t_k $(1 \leq k \leq N_T)$, let the vector $\boldsymbol{\delta}(t_k)$ store the N_Q maximum probabilities

$$\delta_{q_i}(t_k) = \max_{q_j} \Pr[q(t_k) = q_i | q(t_{k-1}) = q_j; O^{(k)}] \,, \quad \text{(A1)}$$

with $1 \leq i \leq N_Q$, and let the vector $\boldsymbol{\Phi}(t_k)$ store the hidden states at t_{k-1} leading to the corresponding maximum probabilities in $\boldsymbol{\delta}(t_k)$, viz.

$$\Phi_{q_i}(t_k) = \arg\max_{q_j} \Pr[q(t_k) = q_i | q(t_{k-1}) = q_j; O^{(k)}] \,, \quad \text{(A2)}$$

with $O^{(k)} = \{o(t_0), \ldots, o(t_k)\}$ and

$$\Pr[q(t_k) = q_i | q(t_{k-1}) = q_j; O^{(k)}] = L_{o(t_k) q_i} A_{q_i q_j} \delta_{q_j}(t_{k-1}) \,. \quad \text{(A3)}$$

The components of $\boldsymbol{\delta}(t_k)$ and $\boldsymbol{\Phi}(t_k)$ are filled by running forward through the N_T observations, then the optimal path $Q^*(O)$ is reconstructed by backtracking.

Note that Higgs & Attwood use notation V_k (stands for Viterbi) for the vector delta

1. *Initialization:*

$$\delta_{q_i}(t_0) = L_{o(t_0)q_i}\Pi_{q_i}, \tag{A4}$$

for $1 \leq i \leq N_Q$.

2. *Recursion:*

$$\delta_{q_i}(t_k) = L_{o(t_k)q_i} \max_{1\leq j\leq N_Q}[A_{q_iq_j}\delta_{q_j}(t_{k-1})], \tag{A5}$$

$$\Phi_{q_i}(t_k) = \arg\max_{1\leq j\leq N_Q}[A_{q_iq_j}\delta_{q_j}(t_{k-1})], \tag{A6}$$

for $1 \leq i \leq N_Q$ and $1 \leq k \leq N_T$.

3. *Termination:*

$$\max \Pr(Q|O) = \max_{q_j} \delta_{q_j}(t_{N_T}) \tag{A7}$$

$$q^*(t_{N_T}) = \arg\max_{q_j} \delta_{q_j}(t_{N_T}) \tag{A8}$$

for $1 \leq j \leq N_Q$.

4. *Optimal path backtracking:*

$$q^*(t_k) = \Phi_{q^*(t_{k+1})}(t_{k+1})$$

for $0 \leq k \leq N_T - 1$.

The language of the protein universe

Andrea Scaiewicz and Michael Levitt

Proteins, the main cell machinery which play a major role in nearly every cellular process, have always been a central focus in biology. We live in the post-genomic era, and inferring information from massive data sets is a steadily growing universal challenge. The increasing availability of fully sequenced genomes can be regarded as the 'Rosetta Stone' of the protein universe, allowing the understanding of genomes and their evolution, just as the original Rosetta Stone allowed Champollion to decipher the ancient Egyptian hieroglyphics. In this review, we consider aspects of the protein domain architectures repertoire that are closely related to those of human languages and aim to provide some insights about the language of proteins.

Address
Department of Structural Biology, Stanford University, Stanford, CA 94305-5126, United States

Corresponding author: Levitt, Michael (michael.levitt@stanford.edu)

Current Opinion in Genetics & Development 2015, **35**:50–56

This review comes from a themed issue on **Genomes and evolution**

Edited by **Antonis Rokas** and **Pamela S Soltis**

For a complete overview see the Issue and the Editorial

Available online 3rd November 2015

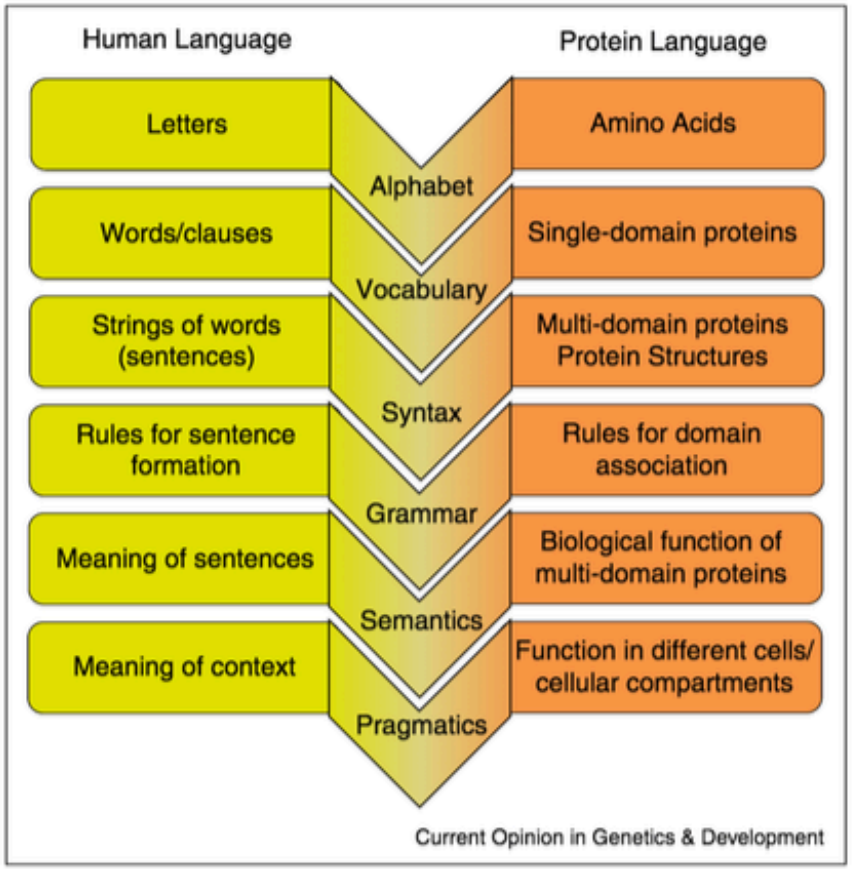
http://dx.doi.org/10.1016/j.gde.2015.06.010

The vocabulary of proteins

Protein domains correspond to the words in the proteins language. Domains can be distinguished by their sequence (sequence profiles) or their structure (classification databases). Sequence profiles are most commonly represented using statistical models such as position specific scoring matrices (PSSM) and hidden Markov models (HMM). HMM-based methods include Pfam [9], EVEREST [10], SMART [11], and PANTHER [12]. PSSM-based methods include PRINTS [13], PROSITE [14] and ProDom [15]. Here we refer to sequence profiles as domains or words. Structure-based classifications including SCOP [16,17], SCOP2 [18] and CATH [19–21], as well as predicted domain structures as in SUPERFAMILY [22,23], Gene3D [24,25], ECOD [26] and COPS [27] are not considered here.

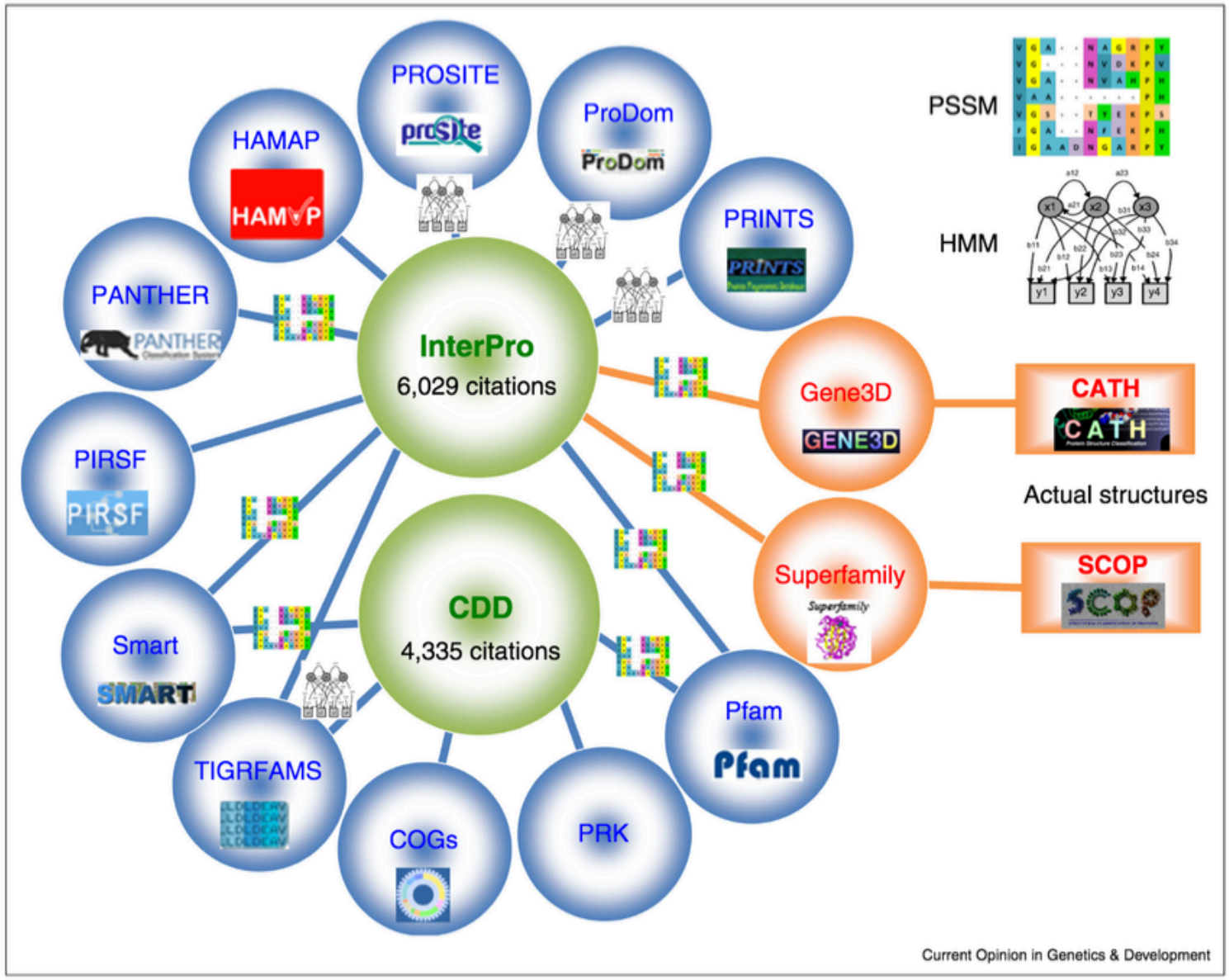
Sequence-based methods differ in coverage, level of curation and definition of families. Compilation databases like CDART [28–30] and InterPro [31] match all sets of profiles to all known sequences (Figure 2). This facilitates the classification of the protein universe into protein families by providing the locations of different domains along every sequence. Often two different sequence profiles match the same region of sequence leading to several domains, or words, for the same physical object. This can lead to confusion and such synonyms need to be recognized and possibly eliminated [32••].

Figure 1



Analogy between human and proteins languages. In this comparison, the vocabulary (domains) of proteins is built from an alphabet of amino acids. The syntax principles enable domain association to form multi-domain architectures, a process governed by hierarchical rules (grammar), that determine the structure and hence the biological function (semantics) of proteins. In several languages, for example in English, a number of different classes of words exist (nouns, adjectives, verbs, adverbs, pronouns, conjunctions). Each class has its task in the language, that is, nouns name words, adjectives describe nouns, verbs are action words, conjunction connect words. Analogously, one can also distinguish different classes of domains with different tasks (motors, binding proteins, enzymes, signaling proteins, structural proteins, targeting proteins).

Figure 2



Sequence profile databases can be sequence-based (blue circles) or structure-based (orange circles). Sequence-based profiles are derived by mainly two methods: HMMs (Hidden Markov Models) or PSSMs. (Position Sensitive Sequence Matrices). Structure-based profiles in Gene3D and superfamily are generated from HMMs built from actual structures coming from CATH and SCOP, respectively. Two main integrative resources, CDART and InterPro, are shown (green circles) with the databases they include.