

## Indice

Sommario.....	5
1 Introduzione.....	6
1.1 Geni e proteine.....	6
1.2 La ricerca genomica e proteomica.....	8
1.3 Vocabolari controllati, ontologie e annotazioni funzionali.....	9
1.3.1 Il concetto di annotazione.....	10
1.4 Banche dati biomolecolari.....	11
1.5 Il Progetto Gene Ontology.....	12
1.6 Organizzazione dell'elaborato.....	16
2 Il problema.....	17
2.1 Difficoltà nell'efficace utilizzo delle informazioni biomolecolari disponibili.....	17
2.2 Motivazioni dell'utilizzo delle tecniche per la predizione.....	18
3 La soluzione proposta.....	19
3.1 Rappresentazione di annotazioni biomolecolari come matrice delle occorrenze.....	19
3.2 Latent Semantic Indexing.....	21
3.3 Singular Values Decomposition.....	21
3.2.1 L'algoritmo.....	25
3.2.2 LSI tramite SVD.....	26
3.2.3 Limiti di LSI.....	28
3.3 L'architettura del software AnnotationPredictor.....	28
3.3.1 Costruzione della matrice delle annotazioni.....	29
3.3.2 Anomalie e loro correzione.....	29

---

3.3.3	Dettagli implementativi .....	31
4	Obiettivi del progetto di tesi .....	32
5	Reingegnerizzazione del software AnnotationPredictor.....	33
5.1	Modifiche puntuali.....	35
6	Implementazione di nuove tecniche per l'analisi semantica.....	37
6.1	Probabilistic Latent Semantic Analysis .....	37
6.1.1	Il concetto di <i>topic</i> .....	37
6.1.2	Il modello formale.....	38
6.1.3	Algoritmo Expectation-Maximization .....	40
6.1.4	pLSAnorm: una variante di pLSA .....	42
6.1.5	Applicazione di pLSA.....	43
6.1.6	Problemi aperti.....	43
6.1.7	Confronto tra LSI e pLSA.....	45
6.2	Schemi di peso .....	48
6.2.1	Pesi locali .....	51
6.2.2	Pesi globali.....	51
6.2.3	Normalizzazione .....	51
6.2.4	Composizione degli schemi .....	52
7	Validazione .....	53
7.1	Categorie di annotazioni .....	53
7.2	Validazione .....	53
7.2.1	Validazione tramite curve ROC.....	54
7.2.2	Validazione su più database.....	56
7.2.3	Validazione nella letteratura .....	57

---

8 Risultati .....	58
8.1 Stima dei parametri del modello pLSA .....	58
8.1.1 Stima della soglia di stop per l'algoritmo EM .....	58
8.1.2 Numero di <i>topics</i> .....	63
8.1.3 Variazioni dovute alla inizializzazione del modello.....	71
8.2 Confronto tra pLSA e LSI.....	73
8.2.1 Confronto basato su curve ROC .....	73
8.2.2 Confronto basato sulla percentuale di predizioni confermate.....	78
8.3 Schemi di peso .....	79
9 Conclusioni .....	83
10 Sviluppi Futuri .....	84
10.1 Sviluppi pLSA .....	84
10.2 Sviluppi nella predizioni di annotazioni .....	84
12 Bibliografia .....	86
Appendice - Datasets .....	88
Indice delle figure .....	91
Indice delle tabelle .....	93