MINISTERO DELL'ISTRUZIONE, DELL'UNIVERSITA' E DELLA RICERCA

DIREZIONE GENERALE RICERCA

PROGETTO DI RICERCA - MODELLO A

BANDO FIRB - PROGRAMMA "FUTURO IN RICERCA"

Anno 2008 - Protocollo: RBFR086NN1

Linea d'intervento 2

1 - Titolo del Progetto di Ricerca

Italiano
Inferenza e ottimizzazione in sistemi complessi: dalla termodinamica dei vetri di spin agli algoritmi di message-passing

Inglese
Inference and optimization in complex systems: from the thermodynamics of spin glasses to message passing algorithms

2 - Durata del Progetto di Ricerca

48 mesi

3 - Coordinatore scientifico della ricerca (Principal Investigator)

RICCI TERSENGHI	***Federico***	***RCCFRC72L09H501Q***
(cognome)	(nome)	(Codice Fiscale)
Ricercatore confermato		***09/07/1972***
(qualifica)		(data di nascita)
Università degli Studi di ROMA "La Sapienza"	***Dipartimento di FISICA***	
(Istituzione di appartenenza)	(Dipartimento/Istituto/Divisione/Settore)	
0649914052	***064957697***	***federico.ricci@roma1.infn.it***
(telefono)	(fax)	(e-mail)

4 - Abstract del Progetto di Ricerca

Italiano
I problemi di inferenza statistica e di ottimizzazione sono forse quelli maggiormente diffusi tra tutte le discipline scientifiche e non solo. Ad esempio l'inferenza bayesiana è di fondamentale importanza in tutti gli ambiti in cui si deve estrarre il massimo delle informazioni dalle evidenze sperimentali. Mentre l'ottimizzazione, ossia la massimizzazione o minimizzazione di una data funzione, è un problema centrale in moltissime applicazioni reali (quali ad esempio tutti i problemi di 'scheduling') e in tutti gli approcci basati sulla massimizzazione della verosimiglianza.

In generale entrambi questi problemi possono essere riscritti nella seguente forma: data una distribuzione di probabilità congiunta di N variabili P(x1,...xN), si calcolino le probabilità marginali su un sottoinsieme delle variabili.
Questo problema è strettamente legato al calcolo dell'energia libera che si fa usualmente in meccanica statistica. Si pensi, ad esempio, che il calcolo della magnetizzazione media un una singola variabile di spin (cioè a due valori) è equivalente al calcolo della probabilità marginale su quella stessa variabile.

Il calcolo esatto delle marginali può essere fatto solo in casi molto particolari ed è quindi di scarso uso nei problemi reali. Nei casi piú generali si devono invece usare dei metodi approssimati (si pensi ad esempio ai metodi Monte Carlo), ma questi sono tipicamente troppo lenti per essere applicati in tutti i casi di interesse.

Esiste tuttavia una categoria di algoritmi molto veloci, noti con il nome di algoritmi di "message-passing", in cui la stima delle probabilità marginali viene eseguita attraverso un processo iterativo: durante questo processo le variabili si scambiano la migliore stima attuale delle loro rispettive marginali, fino all'eventuale raggiungimento di un punto fisso. Il piú noto algoritmo in questa categoria è forse quello di "Belief Propagation" (BP) che ha avuto straordinari successi soprattutto nella teoria dei codici di correzione di errori, permettendo di arrivare molto vicino alla soglia di Shannon (limite superiore per la decodifica dei messaggi corrotti). Dalle probabilità marginali del punto fisso di BP si ottiene il valore dell'energia libera in approssimazione di Bethe-Peierls.

Nonostante BP sia un algoritmo molto veloce, la sua convergenza al punto fisso è assicurata solo sotto condizioni di scarso interesse pratico. Ad esempio, nell'ambito dei problemi di soddisfacimento di vincoli su grafi aleatori (random Constraint Satisfaction Problems, rCSPs), che sono un prototipo di problema computazionale difficile molto studiato nella Theoretical Computer Science, è stato osservato che aumentando la densità dei vincoli la struttura delle soluzioni subisce delle vere e proprie transizioni di fase (per N molto grande). A queste transizioni sono associati dei fenomeni "dinamici" quali la mancata convergenza di BP. In questi casi i metodi tradizionali di inferenza e ottimizzazione sono praticamente inutili.

Recentemente i rCSP sono stati studiati a fondo con tecniche provenienti dalla meccanica statistica dei sistemi disordinati e ciò ha permesso di comprendere l'origine della crescita della complessità computazionale vicino ai punti critici. In particolare è stato proposto un nuovo algoritmo di message-passing noto con il nome di "Survey Propagation" (SP) che riesce a trovare soluzioni anche a densità di vincoli superiori alla soglia limite per la convergenza di BP. Questo è stato un importantissimo risultato, che ha mostrato chiaramente l'efficienza degli algoritmi basati su idee di tipo "fisico". Purtroppo queste tecniche di risoluzione estremamente efficienti sono per il momento relegate al caso di modelli definiti su grafi aleatori, di scarso interesse per le applicazioni.

Riteniamo tuttavia che sia possibile estendere questi veloci algoritmi anche a problemi definiti su grafi con topologia non completamente aleatoria.
Lo scopo di questo progetto è proprio quello di generalizzare gli algoritmi di message-passing (BP e SP) affinché tengano conto dei cicli corti (e di altre strutture topologiche fortemente correlate), spesso presenti nelle reti generate da problemi reali.

Il progetto contiene sia un parte piú teorica in cui useremo le piú potenti tecniche analitiche sviluppate nel campo della meccanica statistica dei sistemi disordinati (metodo delle repliche e della cavità) in unione con altre approssimazioni quali il Cluster Variation Method, sia un parte in cui progetteremo e realizzeremo simulazioni numeriche su grande scala (per le quali chiediamo apposite risorse di calcolo).

I risultati che contiamo di ottenere alle fine di questo progetto sono degli algoritmi molto veloci per risolvere problemi di inferenza ed ottimizzazione su grafi con molte strutture locali, come quelli che tipicamente sono presenti in applicazioni reali.

Questi algoritmi saranno da noi testati su problemi di interesse pratico notoriamente difficili, quali ad esempio quelli contenuti nella libreria SATLIB disponibile su Internet o alcuni problemi di origine biologica (ad esempio reti metaboliche, reti di regolazione trascrizionale e unsupervised data clustering).

Le ricadute di questo progetto possono essere molteplici e molto importanti, perché nuovi e piú potenti algoritmi per l'inferenza e l'ottimizzazione troverebbero immediata applicazione nei piú svariati campi della scienza e della tecnologia. Si pensi, ad esempio, all'ottimizzazione di un processo produttivo o allo 'screening' probabilistico delle malattie; ma soprattutto riteniamo che l'impatto maggiore di questi nuovi algoritmi si noterebbe nella biologia computazionale, un campo in notevole espansione grazie all'aumento vertiginoso dei dati a disposizione.

I partecipanti al progetto sono tutti giovani ricercatori (di 36, 35 e 33 anni) che hanno una solida formazione di base in meccanica statistica dei sistemi disordinati e nelle simulazioni numeriche(lavorano tutti nel gruppo che si è formato intorno alla figura del Prof. Giorgio Parisi). Vantano inoltre un ottima esperienza nel campo in cui si svolge il progetto: hanno infatti già partecipato ad alcuni grandi progetti in questo campo di ricerca a livello europeo dando contributi fondamentali.
Sono infine in stretto contatto con importanti gruppi stranieri (sia teorici che sperimentali) che lavorano nel campo della biologia computazionale.
Non ultimo, il luogo dove si svolgerà la ricerca, il Dipartimento di Fisica della Sapienza di Roma, offre un ambiente di lavoro eccellente ed estremamente stimolante, permettendo l'interazione con molti professori di esperienza e con altrettanti brillanti giovani ricercatori.

In conclusione, sebbene questo progetto sia molto ambizioso, riteniamo che i proponenti abbiamo tutte le carte in regola per portarlo a compimento, ottenendo dei risultati che potrebbero avere sorprendenti ricadute, al di là delle piú positive aspettative.

Inglese
Problems involving statistical inference and optimization are perhaps the most common in science and in many applied fields. For example, Bayesian inference has a central role whenever one has to retrieve information from experimental data; while optimization, namely the maximization or minimization of a prescribed function, is crucial in many real-world applications (like scheduling problems) and in any approach based on the maximization of the likelihood.

In general, both these problems can be recast in the following form: given a joint probability distributions of N variables P(x1,...,xN), compute the marginal probabilities over a subset of the variables.

This problem is strictly linked to the calculation of the free energy that is usually carried out in statistical mechanics. For example, the calculation of the average magnetization of a single spin (or Boolean) variable is equivalent to the calculation of the marginal probability over the same variable.

The exact calculation of marginals can be performed only in very specific cases and is rarely employed in real world applications. In more standard instances, one has to resort to approximate methods (like Monte Carlo schemes). Unfortunately the latter are normally too slow to be of use in all cases of interest.

There exists however a class of fast algorithms, known as message-passing algorithms, through which marginals are computed via an iterative procedure in which variables exchange the best estimates about the respective marginals among themselves, until a fixed point is eventually reached. The best known such algorithm is perhaps the one called Belief Propagation (BP). BP proved to be extremely successful, especially in coding theory and in error correcting codes, allowing to almost saturate Shannon's bound (the limit above which the decoding of corrupted codes is impossible). The calculation of the marginal probabilities at the fixed point of BP is equivalent to the calculation of the free energy in the Bethe-Peierls approximation.

While BP is a vary fast algorithm, its convergence to a fixed point is ensured only in cases of limited practical interest. For example, in random constraint satisfaction problems (RCSPs, prototypical combinatorial optimization problems widely studied in theoretical computer science) it has been observed that increasing the density of constraints the structure of the solutions undergoes various types of phase transitions (for large N). These transitions are in turn related to dynamical phenomena like the failed convergence of BP. In such cases, traditional methods of inference and optimization are practically useless.

Recently RCSPs have been studied in detail employing techniques derived in the statistical mechanics of disordered systems, and this approach has shed some light on the origin of the increase in computational complexity close to the critical points. In particular, a new message-passing algorithm called Survey Propagation (SP) has been proposed, which is able to locate solutions even when the constraint density exceeds that above which BP fails to converge. This important result has clearly shown that "physical" algorithms can be very effective. The downside is that, as of now, such techniques appear to be suited only when the underlying problem is defined on a random graph (as RCSPs), and real world problems rarely come in this form.

Our idea is that it is possible to generalize these fast algorithms to deal with problems whose underlying topology is not completely random. The goal of this project is precisely to modify message-passing algorithms (BP and SP) to take into account the effects induced by short loops and by other strongly correlated topological structures that often characterize real world problems.

The project contains both a theoretical part, where we will use the finest analytical techniques available for the statistical study of disordered systems (the replica method and the cavity method) together with other types of approximations like the Cluster Variation Method, and a strictly computational part where large scale simulations (to carry out which we request specific computational facilities) will be performed.

As a result we expect to obtain improved (very fast) algorithms to solve inference and optimization problems on graphs with various local structures, like those that appear in most applications.

These algorithms will be tested on instances of practical relevance that are known to be particularly difficult, like those contained in the SATLIB libraries or like problems of biological origin such as metabolic networks, transcriptional regulation networks or unsupervised data clustering.

The potential applicative impact of this project is huge, since faster algorithms for inference and optimization would be useful in many different branches of science and technology. One could think, for example, to the optimization of scheduling processes or the probabilistic screening of diseases; above all, we think that computational biology with its growing amount of available data would enormously benefit from the introduction of improved algorithms.

The project participants are young researchers (36, 35 and 33 years old) with a solid background formation in statistical mechanics of disordered systems and in numerical simulations (they all work in the research group grown around Prof. Giorgio Parisi). In addition, they share a considerable experience in research at the international level, as they have all taken part in large European research projects in the field giving fundamental contributions. Finally, they have standing collaborations with major foreign research groups (both theoretical and experimental) working in the field of computational biology. Last but not least, the research will be carried out at the Physics Department of Rome University La Sapienza, a particularly stimulating environment where the interaction with more experienced as well as younger brilliant researchers is possible.

In summary, while this is an ambitious project, we believe that the proponents are fit to face the challenges it presents. In turn the results may bear a surprising impact, even beyond very optimistic expectations.

5 - Parole chiave

Italiano

1. *meccanica statistica*
2. *sistemi disordinati*
3. *inferenza statistica*
4. *ottimizzazione*
5. *cluster variation method*

Inglese

1. *statistical mechanics*
2. *disordered systems*
3. *statistical inference*
4. *optimization*
5. *cluster variation method*

6 - Settori di ricerca ERC (European Research Council) interessati dal Progetto di Ricerca

PE Mathematics, physical sciences, information and communication, engineering, universe and earth sciences

PE3 Condensed matter in physics and chemistry: condensed matter (structure, electronic properties, fluids,…), statistical physics, nanosciences, reactions
PE3_3 Statistical physics

PE1 Mathematical foundations: all areas of mathematics, pure and applied,plus mathematical aspects of theoretical computer science, and mathematical physics
PE1_10 Theoretical computer science

LS Life Sciences

LS2 Genetics, genomics, bioinformatics and systems biology: molecular and cell genetics, genomics, transcriptomics, proteomics, metabolomics, bioinformatics,computational biology, biostatistics, biological modelling and simulation, systems biology
LS2_14 Computational biology

7 - Curriculum scientifico

Italiano
Federico Ricci Tersenghi è nato a Roma il 9/7/1972.

FORMAZIONE E CARRIERA SCIENTIFICA

2002-oggi Ricercatore universitario (SSD FIS/02) nella Facoltà di Scienze MFN della Sapienza di Roma.
2007 Ottiene una posizione di Visiting Assistant Researcher presso la University of California, Berkeley, per un periodo di 7 mesi.
1998-2001 Post-doc presso l'International Center for Theoretical Physics (ICTP) di Trieste.
1995-98 Dottorato di ricerca in Fisica alla Sapienza di Roma sotto la supervisione del Prof. Giorgio Parisi. Titolo della tesi: "Off-equilibrium dynamical studies in disordered systems".
11/1995 Primo classificato nell'esame di ammissione al Corso di Perfezionamento presso la Scuola Normale Superiore di Pisa.
1990-95 Laurea in Fisica alla Sapienza di Roma (voto medio 29,94/30) e diploma con 110 e lode.

PREMI

Nel 1992 e 1994 riceve il Premio in Fisica "E. Persico" assegnato dall'Accademia Nazionale dei Lincei.

PUBBLICAZIONI SCIENTIFICHE

Autore del testo universitario "Programmazione Scientifica" (Pearson Education Italia, 2006) di 650 pagine, pensato come supporto didattico per tre corsi della laurea triennale.

Autore di 56 articoli a carattere scientifico pubblicati su riviste internazionali con referee, che sono stati citati circa 1400 volte, per un valore del cosiddetto H factor pari a 22 (dati estratti da Google Scholar, a causa dell'incompletezza del database ISI).

È stato Guest Editor per lo Special Issue su "Statistical Physics of Disordered Systems: from real materials to optimization and codes", J. Phys. A. 36 (2003).

Svolge regolarmente attività di referee per tutte le piú importanti riviste del settore: PRL, PRB, PRE, JPA, EPL, EPJB, JSTAT.

PRESENTAZIONI ORALI

Dal 2000 è stato chiamato come "invited speaker" a 16 tra scuole e conferenze internazionali.

CONFERENZE INTERNAZIONALI ORGANIZZATE

1) "Statistical Physics and Computational Problems: Beyond the Analogy". Parigi, giugno 2004.
2) "Fundamental Aspects of Complexity". ICTP, settembre 2004.
3) "Statistical Physics of Glasses, Spin Glasses, Information Processing and Combinatorial Optimization". Les Houches, febbraio 2005.
4) "Statistical Physics of Disordered Systems and its Applications". Accademia dei Lincei, settembre 2005.
5) "Statistical Physics of Glasses, Spin Glasses, Information Processing and Combinatorial Optimization". Les Houches, febbraio 2006.
6) "Common Concepts in Statistical Physics and Computer Science". ICTP, luglio 2007.

7) "EPS - CMD 22, 2008. The 22nd General Conference of the Condensed Matter Division of the European Physical Society". Sapienza, agosto 2008.
8) "Wandering with Curiosity in Complex Landscapes: A scientific conference in honor of Giorgio Parisi for his 60th birthday". Sapienza e Accademia dei Lincei, settembre 2008.
9) "ECCS'08. The 5th European Conference on Complex Systems". Gerusalemme, settembre 2008.

PROGETTI FINANZIATI

2002 coordina "Progetto Giovani Ricercatori" (Murst).
2002 partecipa Cofin (Murst).
2002-06 partecipa agli ECC MTR Network DYGLAGEMEM e STIPCO, svolge parte del coordinamento scientifico.
2004-07 partecipa all'ECC Integrated Project EVERGROW, responsabile del WorkPackage su "Belief and survey".
2005 coordina "Progetto di Supercalcolo" (INFM-CINECA), 30k ore di calcolo parallelo.
2006 partecipa PRIN (221kE dal Miur).
2008 coordina "Grandi attrezzature scientifiche" (25kE dalla Sapienza).
dal 2002 partecipa a progetti della Sapienza finanziati ogni anno con circa 20kE.

ATTIVITÀ DIDATTICA

A partire dal 2003 ha tenuto in affidamento i corsi di "Laboratorio di Fisica Computazionale" (I e II, 6 CFU) e di "Calcolo delle Probabilità" (6 CFU) all'interno del corso di laurea in Fisica della Sapienza di Roma.

È stato supervisore di 15 tesi di laurea e di una tesi di dottorato di ricerca.

COMUNICAZIONE SCIENTIFICA

2006 scrive la voce "Simulazioni di processi fisici mediante calcolatore" per l'Enciclopedia Treccani.
dal 2007 co-organizza "Caffè Scienza" con cadenza circa mensile.

Inglese
Federico Ricci-Tersenghi (born July 7, 1972)

EDUCATION AND ACADEMIC CAREER

2002-present Assistant Professor (Ricercatore) at the Physics Department of Sapienza University of Rome.
2007 Invited as a Visiting Assistant Researcher to the University of California at Berkeley for 7 months.
1998-2001 Post-Doc at the International Center for Theoretical Physics (ICTP) in Trieste.
1995-98 Ph.D. student in Physics at Sapienza University of Rome. Thesis on "Off-equilibrium dynamical studies in disordered systems".
11/1995 He ranked first at the admission test for the Ph.D program at Scuola Normale Superiore in Pisa.
1990-95 Graduate studies in Physics at Sapienza University of Rome with an average mark of 29.94/30 and degree Cum Laude.

HONORS

In 1992 and in 1994 he got the Price in Physics "E. Persico" assigned by the Accademia Nazionale dei Lincei.

SCIENTIFIC PUBLICATIONS

Coauthor of the textbook "Programmazione Scientifica" (650 pages, Pearson Education Italia, 2006), a didactic support for 3 courses on Computational Physics.

Author of 56 scientific works published on international journals with referee, which have been cited roughly 1400 times, thus providing a "H factor" equal to 22 (data extracted from Google Scholar, because ISI database is incomplete).

Guest Editor for the Special Issue on "Statistical Physics of Disordered Systems: from real materials to optimization and codes", J. Phys. A 36 (2003).

Acts as a referee for all the most important journal in the field: PRL, PRB, PRE, JPA, EPL, EPJB, JSTAT.

ORAL PRESENTATIONS

Invited speaker to 16 international conferences and schools (since 2000).

CONFERENCES ORGANIZED

1) "Statistical Physics and Computational Problems: Beyond the Analogy". Paris, June 2004.
2) "Fundamental Aspects of Complexity". ICTP, September 2004.
3) "Statistical Physics of Glasses, Spin Glasses, Information Processing and Combinatorial Optimization". Les Houches, February 2005.
4) "Statistical Physics of Disordered Systems and its Applications". Accademia dei Lincei, September 2005.
5) "Statistical Physics of Glasses, Spin Glasses, Information Processing and Combinatorial Optimization". Les Houches, February 2006.
6) "Common Concepts in Statistical Physics and Computer Science". ICTP, July 2007.
7) "EPS - CMD 22, 2008. The 22nd General Conference of the Condensed Matter Division of the European Physical Society". Sapienza, August 2008.
8) "Wandering with Curiosity in Complex Landscapes: A scientific conference in honor of Giorgio Parisi for his 60th birthday". Sapienza and Accademia dei Lincei, September 2008.
9) "ECCS'08. The 5th European Conference on Complex Systems". Jerusalem, September 2008.

GRANTS AWARDED

2002 PI in "Progetto Giovani Ricercatori" (Murst).
2002 participant to Cofin (Murst).
2002-06 participant to ECC MTR Networks STIPCO and DYGLAGEMEM: scientific coordinator support.
2004-07 participant to ECC Integrated Project EVERGROW: manager of workpackage on "Belief and survey".
2005 PI in "Progetto di Supercalcolo" (INFM-CINECA): 30k parallel computing hours awarded.
2006 participant to PRIN (Miur): 221kE awarded.
2008 PI in "Progetto grandi attrezzature scientifiche" (Sapienza): 25kE awarded
since 2002 participant to projects funded yearly by Sapienza University of Rome with roughly 20kE.

DIDACTIC ACTIVITIES

Since 2003 teacher of the courses "Computational Physics" (undergraduate) and "Probability Theory / Stochastic Calculus" (master) in the Physics career at Sapienza University of Rome.

Supervisor of 15 graduate thesis and one Ph.D. thesis of Tommaso Castellani (2003-2006) on "The multi p-spin model: comparing TAP states and out-of-equilibrium dynamics".

SCIENTIFIC COMMUNICATION ACTIVITIES

2006 Author of the entry on "Simulazioni di processi fisici mediante calcolatore" (Computer simulations of physical processes) for the Italian encyclopedia, edited Treccani.
since 2007 Co-organizer of a monthly "Caffè Scienza" (café scientifique) in Rome.

8 - Pubblicazioni scientifiche più significative del Coordinatore della Ricerca

nº	Pubblicazione
1.	*KRZAKALA F, MONTANARI A, RICCI-TERSENGHI F., SEMERJIAN G, ZDEBOROVA L (2007). Gibbs States and the Set of Solutions of Random Constraint Satisfaction Problems. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, vol. 104; p. 10318-10323, ISSN: 0027-8424*
2.	*MONTANARI A, RICCI-TERSENGHI F., SEMERJIAN G (2008). Clusters of solutions and replica symmetry breaking in random k-satisfiability. JOURNAL OF STATISTICAL MECHANICS: THEORY AND EXPERIMENT, vol. -; p. P04004--, ISSN: 1742-5468, doi: 10.1088/1742-5468/2008/04/P04004*
3.	*MONTANARI A, RICCI-TERSENGHI F., SEMERJIAN G (2007). Solving Constraint Satisfaction Problems through Belief Propagation-guided decimation. In: Proceedings of the 45th Allerton Conference on Communication, Control and Computing. University of Illinois, 26-28 settembre 2007, p. 352-359*
4.	*MANNINO C, ORIOLO G, RICCI-TERSENGHI F., CHANDRAN S (2007). The stable set problem and the thinness of a graph. OPERATIONS RESEARCH LETTERS, vol. 35; p. 1-9, ISSN: 0167-6377*
5.	*ACHLIOPTAS D, RICCI-TERSENGHI F. (2006). On the Solution-Space Geometry of Random Constraint Satisfaction Problems. In: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing. Seattle, Washington, USA, 21-23 maggio 2006, NEW YORK: ACM Press, p. 130-139, ISBN/ISSN: 1-59593-134-1*
6.	*BARONE L.M., MARINARI E., ORGANTINI G., RICCI-TERSENGHI F. (2006). PROGRAMMAZIONE SCIENTIFICA. MILANO: Pearson Education Italia s.r.l., p. 1-622, ISBN: 978-8-8719-2242-3*
7.	*JORG T, RICCI-TERSENGHI F. (2008). Entropic Effects in the Very Low Temperature Regime of Diluted Ising Spin Glasses with Discrete Couplings. PHYSICAL REVIEW LETTERS, vol. 100; p. 177203--, ISSN: 0031-9007, doi: 10.1103/PhysRevLett.100.177203*
8.	*LEUZZI L, PARISI G, RICCI-TERSENGHI F., RUIZ-LORENZO J.J (2008). Diluted one-dimensional spin glasses with power law decaying interactions. PHYSICAL REVIEW LETTERS, vol. 101; p. 107203--, ISSN: 0031-9007, doi: 10.1103/PhysRevLett.101.107203*
9.	*MARSILI M., MULET R., RICCI-TERSENGHI F. (2003). Learning to compete and coordinate in a complex world. In: COWAN R.; JONARD N.. Heterogenous agents, interactions and economic performance. vol. 521, p. 61, BERLIN HEIDELBERG: Springer-Verlag, ISBN/ISSN: 3-540-44057-7*
10.	*PARISI G., RICCI-TERSENGHI F., RUIZ-LORENZO JJ. (1996). Equilibrium and off-equilibrium simulations of the 4d Gaussian spin glass. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 29; p. 7943-7957, ISSN: 0305-4470*
11.	*PARISI G., RANIERI P., RICCI-TERSENGHI F., RUIZ-LORENZO JJ. (1997). Mean field dynamical exponents in finite-dimensional Ising spin glass. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 30; p. 7115-7131, ISSN: 0305-4470*
12.	*MARINARI E., PARISI G., RICCI-TERSENGHI F., RUIZ-LORENZO JJ. (1998). Violation of the fluctuation-dissipation theorem in finite-dimensional spin glasses. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 31; p. 2611-2620, ISSN: 0305-4470*
13.	*MARINARI E., PARISI G., RICCI-TERSENGHI F., RUIZ-LORENZO JJ. (1998). Small window overlaps are effective probes of replica symmetry breaking in three-dimensional spin glasses. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 31; p. L481-L487, ISSN: 0305-4470*
14.	*PARISI G., RICCI-TERSENGHI F. (2000). On the origin of ultrametricity. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 33; p. 113-129, ISSN: 0305-4470*
15.	*MARINARI E., PARISI G., RICCI-TERSENGHI F., RUIZ-LORENZO JJ. (2000). Off-equilibrium dynamics at very low temperatures in three-dimensional spin glasses. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 33; p. 2373-2382, ISSN: 0305-4470*
16.	*RICCI-TERSENGHI F., RITORT F. (2000). Absence of ageing in the remanent magnetization in Migdal-Kadanoff spin glasses. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 33; p. 3727-3734, ISSN: 0305-4470*
17.	*SIMON P., RICCI-TERSENGHI F. (2000). Coupled Ising models with disorder. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 33; p. 5985-5991, ISSN: 0305-4470*
18.	*MARINARI E., PARISI G., RICCI-TERSENGHI F., ZULIANI F. (2001). The use of optimized Monte Carlo methods for studying spin glasses. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 34; p. 383-390, ISSN: 0305-4470*
19.	*LEONE M., RICCI-TERSENGHI F., ZECCHINA R. (2001). Phase coexistence and finite-size scaling in random combinatorial problems. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 34; p. 4615-4626, ISSN: 0305-4470*
20.	*PARISI G, RICCI-TERSENGHI F., RUIZ-LORENZO JJ (1998). Dynamics of the four-dimensional spin glass in a magnetic field. PHYSICAL REVIEW. B, CONDENSED MATTER AND MATERIALS PHYSICS, vol. 57; p. 13617-13623, ISSN: 1098-0121*
21.	*PARISI G., RICCI-TERSENGHI F., RUIZ-LORENZO JJ. (1999). Universality in the off-equilibrium critical dynamics of the three-dimensional diluted Ising model. PHYSICAL REVIEW E, vol. 60; p. 5198-5201, ISSN: 1063-651X*
22.	*FRANZ S., RICCI-TERSENGHI F. (2000). Ultrametricity in three-dimensional Edwards-Anderson spin glasses. PHYSICAL REVIEW E, vol. 61; p. 1121-1124, ISSN: 1063-651X*
23.	*PAGNANI A., PARISI G., RICCI-TERSENGHI F. (2000). Glassy Transition in a Disordered Model for the RNA Secondary Structure. PHYSICAL REVIEW LETTERS, vol. 84; p. 2026-2029, ISSN: 0031-9007*
24.	*RICCI-TERSENGHI F., STARIOLO DA., ARENZON JJ. (2000). Two Time Scales and Violation of the Fluctuation-Dissipation Theorem in a Finite Dimensional Model for Structural Glasses. PHYSICAL REVIEW LETTERS, vol. 84; p. 4473-4476, ISSN: 0031-9007*
25.	*ARENZON JJ., RICCI-TERSENGHI F., STARIOLO DA. (2000). Dynamics of the frustrated Ising lattice gas. PHYSICAL REVIEW E, vol. 62; p. 5978-5985, ISSN: 1063-651X*
26.	*RICCI-TERSENGHI F., ZECCHINA R. (2000). Glassy dynamics near zero temperature. PHYSICAL REVIEW E, vol. 62; p. R7567-R7570, ISSN: 1063-651X*
27.	*RICCI-TERSENGHI F., WEIGT M, ZECCHINA R (2001). Simplest random K-satisfiability problem. PHYSICAL REVIEW E, STATISTICAL, NONLINEAR, AND SOFT MATTER PHYSICS, vol. 63; p. 026702, ISSN: 1539-3755*

28.	*PICCO M, RICCI-TERSENGHI F., RITORT F (2001). Chaotic, memory, and cooling rate effects in spin glasses: Evaluation of the Edwards-Anderson model. PHYSICAL REVIEW. B, CONDENSED MATTER AND MATERIALS PHYSICS, vol. 63; p. 174412, ISSN: 1098-0121*
29.	*PAGNANI A., PARISI G., RICCI-TERSENGHI F. (2001). Pagnani, Parisi, and Ricci-Tersenghi Reply. PHYSICAL REVIEW LETTERS, vol. 86; p. 1383, ISSN: 0031-9007*
30.	*RICCI-TERSENGHI F., PARISI G., STARIOLO DA., ARENZON JJ. (2001). Ricci-Tersenghi et al. Reply. PHYSICAL REVIEW LETTERS, vol. 86; p. 4717, ISSN: 0031-9007*
31.	*FRANZ S., LEONE M., RICCI-TERSENGHI F., ZECCHINA R. (2001). Exact Solutions for Diluted Spin Glasses and Optimization Problems. PHYSICAL REVIEW LETTERS, vol. 87; p. 127209, ISSN: 0031-9007*
32.	*MARSILI M., MULET R., RICCI-TERSENGHI F., ZECCHINA R. (2001). Learning to Coordinate in a Complex and Nonstationary World. PHYSICAL REVIEW LETTERS, vol. 87; p. 208701, ISSN: 0031-9007*
33.	*MARINARI E, PAGNANI A, RICCI-TERSENGHI F. (2002). Zero-temperature properties of RNA secondary structures. PHYSICAL REVIEW E, STATISTICAL, NONLINEAR, AND SOFT MATTER PHYSICS, vol. 65; p. 041919, ISSN: 1539-3755*
34.	*BARTHEL W., HARTMANN AK., LEONE M., RICCI-TERSENGHI F., WEIGT M., ZECCHINA R. (2002). Hiding Solutions in Random Satisfiability Problems: A Statistical Mechanics Approach. PHYSICAL REVIEW LETTERS, vol. 88; p. 188701, ISSN: 0031-9007*
35.	*PARISI G., RICCI-TERSENGHI F., RUIZ-LORENZO JJ. (1999). Generalized off-equilibrium fluctuation-dissipation relations in random Ising systems. THE EUROPEAN PHYSICAL JOURNAL. B, CONDENSED MATTER PHYSICS, vol. 11; p. 317-325, ISSN: 1434-6028*
36.	*PICCO M., RICCI-TERSENGHI F., RITORT F. (2001). Aging effects and dynamic scaling in the 3D Edwards-Anderson spin glasses: a comparison with experiments. THE EUROPEAN PHYSICAL JOURNAL. B, CONDENSED MATTER PHYSICS, vol. 21; p. 211-217, ISSN: 1434-6028*
37.	*FRANZ S., MEZARD M., RICCI-TERSENGHI F., WEIGT M., ZECCHINA R. (2001). A ferromagnet with a glass transition. EUROPHYSICS LETTERS, vol. 55; p. 465-471, ISSN: 0295-5075*
38.	*MARINARI E., PARISI G., RICCI-TERSENGHI F., RUIZ-LORENZO JJ., ZULIANI F. (2000). Replica Symmetry Breaking in Short-Range Spin Glasses: Theoretical Foundations and Numerical Evidences. JOURNAL OF STATISTICAL PHYSICS, vol. 98; p. 973-1074, ISSN: 0022-4715*
39.	*HARTMANN A.K, RICCI-TERSENGHI F. (2002). Direct sampling of complex landscapes at low temperatures: the three-dimensional +/-J Ising spin glass. PHYSICAL REVIEW. B, CONDENSED MATTER AND MATERIALS PHYSICS, vol. 66; p. 224419, ISSN: 1098-0121*
40.	*BRAUNSTEIN A., LEONE M., RICCI-TERSENGHI F., ZECCHINA R. (2002). Complexity transitions in global algorithms for sparse linear systems over finite fields. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 35; p. 7559, ISSN: 0305-4470*
41.	*FRANZ S, LEONE M, MONTANARI A, RICCI-TERSENGHI F. (2002). The Dynamic Phase Transition for Decoding Algorithms. PHYSICAL REVIEW E, STATISTICAL, NONLINEAR, AND SOFT MATTER PHYSICS, vol. 66; p. 046120, ISSN: 1539-3755*
42.	*MEZARD M., RICCI-TERSENGHI F., ZECCHINA R. (2003). Two solutions to diluted p-spin models and XORSAT problems. JOURNAL OF STATISTICAL PHYSICS, vol. 111; p. 505, ISSN: 0022-4715*
43.	*MONTANARI A., RICCI-TERSENGHI F. (2003). A microscopic description of the aging dynamics: fluctuation-dissipation relations, effective temperature and heterogeneities. PHYSICAL REVIEW LETTERS, vol. 90; p. 017203, ISSN: 0031-9007*
44.	*MONTANARI A., RICCI-TERSENGHI F. (2003). On the nature of the low-temperature phase in discontinuous mean-field spin glasses. PHYSICAL REVIEW. B, CONDENSED MATTER AND MATERIALS PHYSICS, vol. 33; p. 339, ISSN: 1098-0121*
45.	*MONTANARI A, RICCI-TERSENGHI F. (2003). Aging dynamics of heterogeneous spin models. PHYSICAL REVIEW. B, CONDENSED MATTER AND MATERIALS PHYSICS, vol. 68; p. 224429, ISSN: 1098-0121*
46.	*CASTELLANI T., NAPOLANO V., RICCI-TERSENGHI F., ZECCHINA R. (2003). Bicoloring Random Hypergraphs. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 36; p. 11037, ISSN: 0305-4470*
47.	*RICCI-TERSENGHI F. (2003). Measuring the fluctuation-dissipation ratio in glassy systems with no perturbing field. PHYSICAL REVIEW E, STATISTICAL, NONLINEAR, AND SOFT MATTER PHYSICS, vol. 68; p. 065104, ISSN: 1539-3755*
48.	*MONTANARI A, PARISI G, RICCI-TERSENGHI F. (2004). Instability of one-step replica-symmetry-broken phase in satisfiability problems. JOURNAL OF PHYSICS. A, MATHEMATICAL AND GENERAL, vol. 37; p. 2073-2091, ISSN: 0305-4470, doi: 10.1088/0305-4470/37/6/008*
49.	*GODRECHE C, KRZAKALA F, RICCI-TERSENGHI F. (2004). Non-equilibrium critical dynamics of the ferromagnetic Ising model with Kawasaki dynamics. JOURNAL OF STATISTICAL MECHANICS: THEORY AND EXPERIMENT, vol. 04; p. P04007:1-P04007:22, ISSN: 1742-5468, doi: 10.1088/1742-5468/2004/04/P04007*
50.	*MONTANARI A, RICCI-TERSENGHI F. (2004). Cooling-schedule dependence of the dynamics of mean-field glasses. PHYSICAL REVIEW. B, CONDENSED MATTER AND MATERIALS PHYSICS, vol. 70; p. 134406:1-134406:8, ISSN: 1098-0121*
51.	*CASTELLANI T., KRZAKALA F., RICCI-TERSENGHI F. (2005). Spin glass models with ferromagnetically biased couplings on the Bethe lattice: analytic solutions and numerical simulations. THE EUROPEAN PHYSICAL JOURNAL. B, CONDENSED MATTER PHYSICS, vol. 47; p. 99-108, ISSN: 1434-6028, doi: 10.1140/epjb/e2005-00293-1*
52.	*MAIORANO A, MARINARI E, RICCI-TERSENGHI F. (2005). Edwards-Anderson spin glasses undergo simple cumulative aging. PHYSICAL REVIEW. B, CONDENSED MATTER AND MATERIALS PHYSICS, vol. 72; p. 104411:1-104411:5, ISSN: 1098-0121, doi: 10.1103/PhysRevB.72.104411*
53.	*KRZAKALA F, RICCI-TERSENGHI F. (2006). Aging, memory and rejuvenation: some lessons from simple models. JOURNAL OF PHYSICS: CONFERENCE SERIES, vol. 40; p. 42-49, ISSN: 1742-6588, doi: 10.1088/1742-6596/40/1/005*
54.	*RICCI-TERSENGHI F., MINNECI F, SOLA E, CHERUBINI E, MAGGI L (2006). Multivesicular Release at Developing Schaffer Collateral CA1 Synapses: An Analytic Approach to Describe Experimental Data. JOURNAL OF NEUROPHYSIOLOGY, vol. 96; p. 15-26, ISSN: 0022-3077, doi: 10.1152/jn.01202.2005*
55.	*CAPONE B, CASTELLANI T, GIARDINA I, RICCI-TERSENGHI F. (2006). Off-equilibrium confined dynamics in a glassy system with level-crossing states. PHYSICAL REVIEW. B, CONDENSED MATTER AND MATERIALS PHYSICS, vol. 74; p. 144301:1-144301:7, ISSN: 1098-0121, doi: 10.1103/PhysRevB.74.144301*
56.	*FRANZ S, PARISI G, RICCI-TERSENGHI F. (2008). Mosaic length and finite interaction-range effects in a one dimensional random energy model. JOURNAL OF PHYSICS. A, MATHEMATICAL AND THEORETICAL, vol. 41; p. 324011--, ISSN: 1751-8113, doi: 10.1088/1751-8113/41/32/324011*

9 - Elenco delle Unità di Ricerca (UR)

Unita'	Responsabile dell'Unità di Ricerca	Qualifica	Istituzione di appartenenza	Dipartimento/Istituto/Divisione/Settore	Mesi/ uomo
1	*RICCI TERSENGHI Federico*	*Ricercatore confermato tempo pieno*	*Università degli Studi di ROMA "La Sapienza"*	*FISICA*	*130*

10 - Breve descrizione delle Unità di Ricerca

Italiano
In questo progetto è prevista una unica Unità di Ricerca, localizzata presso il Dipartimento di Fisica dell'Università Sapienza di Roma.
Il coordinatore scientifico del progetto, Dr. Federico Ricci-Tersenghi, è anche il responsabile dell'Unità di Ricerca della Sapienza.
Aderiscono a questa unità anche il Dr. Andrea De Martino (di 35 anni) che è ricercatore a tempo determinato presso l'Istituto Nazionale di Fisica della Materia (INFM) del CNR e il Dr. Tommaso Rizzo (di 33 anni) che è borsista presso il Museo storico della fisica e centro studi e ricerche "E. Fermi" di Roma.
I tre membri dell'unità collaborano già attivamente, perché fanno parte del gruppo di ricerca sui sistemi disordinati che si è sviluppato presso il Dipartimento di Fisica intorno alla figura del Prof. Giorgio Parisi.

*** I COMPONENTI DELL'UNITÀ DI RICERCA ***

L'esperienza scientifica dei tre membri, insieme all'ambiente di lavoro offerto dal Dipartimento di Fisica, sono ideali per poter portare a compimento il presente progetto.

Federico Ricci-Tersenghi vanta un'ottima esperienza nel campo delle simulazioni numeriche di modelli di meccanica statistica: fin dai tempi della sua tesi di laurea (1995) ha progettato e realizzato importanti studi numerici di modelli con disordine, usando potenti computer paralleli e le piú moderne tecniche di simulazione. Queste esperienze saranno fondamentali per quelle parti del progetto in cui non è possibile usare un approccio analitico e si è costretti a sviluppare veloci algoritmi e simulazioni su grande scala.

Inoltre a partire dal 2000 Ricci-Tersenghi ha lavorato molto attivamente sul campo di ricerca in cui si inquadra questo progetto, diventandone un esperto a livello mondiale. Ad esempio, nell'Integrated Project EVERGROW, finanziato dalla Commissione Europea dal 2004 al 2007, ha svolto il ruolo di responsabile del settore di ricerca (Workpackage) su "Belief and survey": i risultati raggiunti all'interno di questo workpackage sono stati definiti dal review panel di EVERGROW letteralmente "world-leading".

Ultimo, ma non meno importante, il coordinatore di questo progetto ha già dimostrato negli ultimi anni di saper portare avanti ricerca di alto livello in modo autonomo o tramite collaborazione internazionali con altri giovani ricercatori: si vedano ad esempio le pubblicazioni scientifiche firmate con colleghi piú giovani di lui, quali A. Montanari, G. Semerjian, F. Krzakala, L. Zdeborova e T. Jorg.

Andrea De Martino è ricercatore a tempo determinato al Centro SMC (Statistical Mechanics and Complexity) del CNR-INFM. Ha ottenuto il dottorato di ricerca presso la SISSA di Trieste e ha lavorato come post-doc presso l'Hahn-Meitner-Institut di Berlino prima di arrivare a SMC nel 2003. È autore di oltre 30 lavori su riviste internazionali.

Dal 2006 la sua ricerca si svolge nell'ambito della biofisica teorica e in particolare nello sviluppo di algoritmi efficaci per l'analisi strutturale di reti biologiche e per la predizione della risposta cellulare a perturbazioni ambientali o genetiche (knock-outs). Piú di recente ha contribuito in modo determinante allo sviluppo di un modello dell'attività metabolica cellulare che riproduce correttamente il comportamento del batterio E.coli in diversi ambienti e consente la predizione teorica dell'essenzialità dei geni in varie condizioni. Collabora con laboratori di biologia teorica e sperimentale negli USA e con gruppi all'ICTP di Trieste, al King's College London e alla Ecole Normale Superieure di Parigi. Si veda http://oldboy.phys.uniroma1.it/, pagina "Metabolic Networks" per una descrizione piú dettagliata della sua attività di ricerca.

Tommaso Rizzo fin dalla tesi di laurea (1999) ha svolto attività di ricerca nel campo della fisica dei sistemi disordinati e ha pubblicato circa trenta lavori su riviste specializzate.

Nel corso della sua carriera ha acquisito una solida padronanza della teoria e dei sofisticati strumenti analitici sviluppati in questo settore, in particolare il metodo delle repliche, il metodo della cavità, l'approccio supersimmetrico e la teoria dei campi replicata.
Affianca a questo bagaglio analitico una peculiare esperienza nell'utilizzo di strumenti informatici per il calcolo simbolico su grande scala, risorsa complementare alle notevoli possibilità di cui il progetto dispone nel campo delle simulazioni numeriche.

A partire dal 2004 ha trascorso due anni presso L'Ecole Normale Superieure di Parigi collaborando al progetto EVERGROW ed estendendo i suoi interessi in direzione applicativa.
In particolare ha lavorato all'utilizzo di tecniche delle fisica statistica per progettare efficienti algoritmi per una vasta classe di applicazioni, che è uno dei più recenti sviluppi del settore ed un aspetto centrale del presente progetto.

In questo ambito ha ideato l'algoritmo basato sull'approssimazione di Bethe che è alla base della piattaforma di misurazione di internet ETOMIC completando una delle milestone del progetto EVERGROW.
Ha inoltre introdotto una nuova classe di approssimazioni e di algoritmi pensati per tenere conto in maniera sistematica dei cicli sui modelli grafici.
Le proprietà ed applicazioni di questi algoritmi sono state studiate in collaborazione con H.J. Kappen, il cui gruppo è all'avanguardia nel campo dell'inferenza Bayesiana su sistemi biologici.

In aggiunta ai tre ricercatori di cui sopra, l'unità verrà dotata di due ricercatori post-doc con contratti a gravare sul progetto stesso. Si tratta di contratti triennali, che contiamo di far partire il primo e il secondo anno, in modo che vi sia sempre un ricercatore a tempo pieno sul progetto durante l'intera durata dello stesso. Nei due anni centrali del progetto ci sarebbero così ben due post-doc a lavorarci a tempo pieno e pensiamo che questo sia molto importante. Infatti il progetto, che è molto ambizioso, è stato suddiviso in 5 sotto-progetti; alcuni di questi sono propedeutici ad altri ed è quindi fondamentale riuscire ad ottenere decisivi risultati alla fine del secondo a terzo anno, per poter raggiungere tutti gli obiettivi entro la fine del progetto.

Nel ruolo di ricercatori post-doc pianifichiamo di assumere con contratto co.co.co. due giovani ricercatori che abbiano conseguito il dottorato di ricerca (e quindi abbiano almeno 3/4 anni di esperienza). Il contratto di riferimento è quello di una azione Marie Curie per "experienced researcher" con compenso lordo annuo di circa 50000 euro. Riteniamo che solo un contratto con queste caratteristiche sia competitivo al livello internazionale e ci permetterà di far partecipare a questo progetto i migliori giovani ricercatori.

*** LA STRUTTURA IN CUI OPERA L'UNITÀ DI RICERCA ***

Oltre al personale dedicato al progetto, è utile ricordare le infrastrutture e l'ambiente di lavoro messo a disposizione dal Dipartimento di Fisica della Sapienza, che sono un vero e proprio valore aggiunto per l'unità di ricerca.

Il Dipartimento di Fisica della Sapienza è il piú grande dipartimento di fisica d'Italia e quello che vanta forse la piú gloriosa storia, a partire dai tempi di Fermi. Oggi, sui suoi 19000 metri quadrati, lavorano in attività di ricerca circa 600 persone: di queste circa 150 sono dipendenti universitari (professori ordinari, associati e ricercatori), circa 100 sono dipendenti dell'Istituto Nazionale di Fisica Nucleare (INFN) e circa 350 hanno posizioni a tempo determinato (studenti del dottorato di ricerca, post-doc, professori visitatori, etc...).
La maggior parte di queste persone hanno intense collaborazioni internazionali, che rendono il Dipartimento un centro di ricerca estremamente attivo e vivace.

L'alta qualità della ricerca che viene svolta nel Dipartimento è testimoniata dall'elevato numero di prestigiosi premi internazionali assegnati ad alcuni dei suoi membri: la medaglia Dirac ai Proff. Parisi (1999) e Maiani (2007), il premio Balzan al Prof. De Bernardis (2006), la medaglia Boltzmann ai Proff. Parisi (1992) e Gallavotti (2007), il premio Feltrinelli ai Proff. Parisi (1986), De Bernardis (2001) e Jona-Lasinio (2006) e il premio Dan David al Prof. De Bernerdis (2009) giusto per citare i piú importanti.
Inoltre alcuni membri del Dipartimento sono o sono stati a capo di importanti istituzioni internazionali: il Prof. Cabibbo è presidente della Pontificia Accademia delle Scienze, il Prof. Maiani è stato direttore del CERN di Ginevra e il Prof. Virasoro è stato direttore dell'ICTP di Trieste.

In particolare il Dipartimento di Fisica della Sapienza vanta una tradizione forse unica al mondo nel campo in cui si inquadra questo progetto, ossia la meccanica statistica dei sistemi disordinati. Molti brillanti ricercatori si sono formati presso questo Dipartimento sotto la guida del Prof. Giorgio Parisi e nel Dipartimento

esiste da sempre un'intensa attività di ricerca di altissimo livello su questi temi.
Questo Dipartimento è certamente il posto migliore in cui potrebbe vedere la luce un progetto come quello qui proposto.

Il Dipartimento ospita anche due 'centri di ricerca e sviluppo' dell'Istituto Nazionale di Fisica della Materia (ora CNR-INFM). Di questi due, il centro su Statistical Mechanics and Complexity (SMC), a cui De Martino appartiene e Ricci-Tersenghi e Rizzo sono associati, offre un ambiente particolarmente stimolante per questo progetto: infatti intorno ad SMC si è creato un gruppo di ricerca fatto interamente di giovani (circa una dozzina), che potrà svolgere un ruolo importante al momento di discutere pubblicamente i risultati ottenuti in questo progetto.

Infine il Dipartimento, ai cui corsi di laurea si immatricolano circa 300 studenti l'anno, offre tre programmi di dottorato di ricerca (Fisica, Astrofisica e Scienza dei Materiali) e coordina la scuola di dottorato Volterra che include altre discipline scientifiche. Questa rappresenta un'ulteriore opportunità per il presente progetto, che potrebbe facilmente attirare giovani interessati a svolgere una tesi di dottorato sulle linee di ricerca descritte in questo progetto.

Il Dipartimento di Fisica della Sapienza è ovviamente disposto a sostenere questo progetto con le adeguate strutture necessarie, ad esempio gli spazi per i ricercatori post-doc e quelli per i nuovi computer da acquisire sui fondi di questo progetto.

*** I COMPITI DELL'UNITÀ DI RICERCA ****

Essendo l'intero progetto basato sul lavoro dell'unità di ricerca qui descritta, i compiti di questa unità coincidono esattamente con quello che è descritto nella sezione "Descrizione della Ricerca" nel modello A e non ci sembra utile copiarlo nuovamente in questa sezione.

Approfittiamo di questo spazio per descrivere la suddivisione dei compiti tra i componenti dell'unità di ricerca.

Federico Ricci-Tersenghi, dato il suo ruolo di coordinatore, parteciperà a tutti i sotto-progetti.
Tuttavia si occuperà in modo diretto soprattutto di quelle parti del progetto con un forte carattere numerico-computazionale (SP2 e SP3).

Andrea De Martino, grazie alla sua esperienza nel campo della biologia computazionale, sarà il principale responsabile di SP4.

Tommaso Rizzo si occuperà maggiormente degli aspetti piú analitici e darà il suo maggior contributo in SP1.

Riguardo i due ricercatori post-doc che contiamo di assumere, il primo svolgerà la propria attività nei primi 3 anni del progetto e si dedicherà soprattutto allo sviluppo degli algoritmi lavorando in SP1 e SP2; mentre il secondo, che si unirà al progetto presumibilmente all'inizio del secondo anno, si dedicherà maggiormente all'applicazione degli algoritmi a problemi estratti da librerie pubbliche e di origine biologica, quindi contribuendo soprattutto a SP3 e SP4.

Tutti i partecipanti al progetto daranno un loro contributo a SP5.

Ci teniamo a sottolineare che, date le dimensioni ridotte dell'unità di ricerca, tutti i ricercatori coinvolti in questo progetto parteciperanno a delle riunioni congiunte con cadenza almeno settimanale per aggiornarsi reciprocamente sullo stato della ricerca e discutere insieme gli eventuali problemi incontrati.

Inglese
This project is made by only one research unit, located at the Physics Department of Sapienza University of Rome.
The Principal Investigator of the project, Dr. Federico Ricci-Tersenghi, is also the coordinator of the Sapienza unit.
Are part of this unit also Dr. Andrea De Martino (35 years old), researcher at the Istituto Nazionale di Fisica della Materia (INFM-CNR), and Dr. Tommaso Rizzo (33 years old), post-doc at Museo storico della fisica e centro studi e ricerche "E. Fermi" in Rome.
The three members of this unit are already collaborating actively among them, since they come from the same research group on disordered systems, that grew up in the Sapienza Physics Department around the figure of Prof. Giorgio Parisi.

*** ABOUT THE MEMBERS OF THE RESEARCH UNIT ****

The scientific expertise of the three members of the research unit and the unique environment offered by the Sapienza Physics Department are in our opinion perfect to bring this project to the success.

Federico Ricci-Tersenghi is a recognized expert in the field of numerical simulations of statistical mechanical models: since the times of his undergraduate thesis (1995) he has designed and accomplished many important numerical studies of disordered models, by using powerful parallel computers and the most advanced numerical techniques. This experience is strictly required for all the parts of the project where an analytic approach is not possible and we will be asked to develop fast algorithms and large scale simulations.

Moreover, starting from year 2000, Ricci-Tersenghi has dedicated most of his time working on the research field to which this project belong, becoming an expert at the international level. Just to report an illustrative example, in the Integrated Project EVERGROW, funded by the European Commission between 2004 and 2007, he played the role of coordinator for the workpackage on "Belief and survey": the results reached in this workpackage have been defined by the EVERGROW review panel, literally, "world-leading".

Last, but not least, Ricci-Tersenghi has already shown in recent years the ability of producing top level research in an independent way and through international collaboration with other young researchers (see e.g. scientific publications coauthored with younger colleagues as A. Montanari, G. Semerjian, F. Krzakala, L. Zdeborova and T. Jorg).

Andrea De Martino is Researcher at the CNR/INFM Centre for Statistical Mechanics and Complexity (SMC) in Rome. He obtained his Ph.D. at SISSA (Trieste) and subsequently worked in Germany as a postdoctoral fellow before joining SMC. He has authored or coauthored over 30 articles in international refereed journals.

His current research (since 2006) focuses on theoretical biophysics and particularly on the development of effective algorithms for the structural analysis of biological networks and for predicting cellular responses to external or internal (knock-outs) perturbations. More recently he has given a new contribution to the development of a model of cellular metabolism that reproduces correctly the behavior of the bacterium E.coli in different environments and allows for a theoretical prediction of gene essentiality in various conditions. He has standing collaborations with theoretical and experimental biology labs in the USA as well as with groups at the ICTP in Trieste, King's College London and Ecole Normale Superieure in Paris.
See http://oldboy.phys.uniroma1.it/, page on Metabolic Networks for a more detailed description of his recent activity.

Tommaso Rizzo has been doing research on the physics of disordered systems since taking his university degree (1999) and has published around thirty papers.

During the course of his scientific career he has acquired a solid mastery of the theory and of the complex analytical tools developed in the field, in particular the replica method, the cavity method, the supersymmetric approach and the replica field theory.
Besides this analytical fund he has a distintive proficiency in the use of programming languages for large-scale computer-assisted symbolic calculus that is complementary to the remarkable resources of the project in the field of numerical simulations.

Starting in June 2004 he spent two years at the Ecole Normale Superieure in Paris, joining the EVERGROW project and extending his interests in an application-oriented direction.
In particular he worked on the use of statistical physics techniques for the design of efficient algorithms for a large class of applications, that is one of the most recent trends in the field and a central feature of the present project.

In this context he designed a Bethe-approximation based algorithm that is at the core of the ETOMIC infrastructure for dynamical measurements of the Internet, fulfilling one of the milestones of the EVERGROW project.
He also introduced a new class of approximations and algorithms in order to take care of loops on graphical models in a systematic way. The scope of these algorithms have been further studied in collaboration with H. J. Kappen, whose group is in the forefront of Bayesian inference on biological systems.

In addition to the three researchers above, the unit will hire two post-docs with a 3-years contract. Most probably these contracts will start on the first and second year of the project, such as to have during the entire project at least one post-doc fully dedicated to the project. During the second and third years we believe it is very important to have 2 post-doc working in parallel: the project is very ambitious and is divided in 5 sub-projects; in some of these sub-projects we have to achieve definite results at the end of the second and third year in order to satisfy all the promised goals for the end of the project.

As a post-doc we plan to hire two young researchers, that already got the Ph.D. degree (and thus having at least 3/4 years of research experience). The salary we have to compare with is that of a Marie Curie action for an 'experienced researcher', i.e. a gross amount per year of roughly 50000 euros. If a contract is under this standard value, it will not be competitive at the international level and we will be unable to hire the best young researchers.

*** THE INFRASTRUCTURES AROUND THE RESEARCH UNIT ***

After having described the personnel which will be dedicated to the project, it is worth reminding why the infrastructures and the working environment offered by the Physics Department in Sapienza University of Rome are extremely valuable.

The Sapienza Physics Department (SPD) is the largest physics department in Italy. It has a long tradition starting from Enrico Fermi. Today, on its 19000 square meters, roughly 600 people work on research activities: of these roughly 150 are professors (full, associate and assistant) of the department, roughly 100 are researchers of the Istituto Nazionale di Fisica Nucleare (INFN) and 350 have temporary positions (Ph.D. students, post-doc, visiting professors, etc...). The SPD has very strong international collaborations, which make it a very active research center.

The high quality of the research carried on in this institution is witnessed by the international prices won by professors of this department: the Dirac medal to Prof. Parisi (1999) and Prof. Maiani (2007), the Balzan prize to Prof. De Bernardis (2006), the Boltzmann medal to Prof. Parisi (1992) and Prof. Gallavotti (2007), the Feltrinelli prize to Prof. Parisi (1986), Prof. De Bernardis (2001) and Prof. Jona-Lasinio (2006), the Dan David prize to Prof. De Bernerdis (2009) just citing most important ones.
Moreover some members of the SPD are or were at the head of important international institutions: Prof. Maiani, former director of CERN, Prof. Virasoro, former director of ICTP and Prof. Cabibbo, president of the Pontifical Academy of Science.

The SPD is hosting two "centri di ricerca e sviluppo" (research and development centers) of the Istituto Nazionale di Fisica della Materia (CNR-INFM). One of the two is the center SMC (Statistical Mechanics and Complexity) which De Martino is part of and Ricci-Tersenghi and Rizzo are associated to. SMC offers a particularly stimulating environment for this project: roughly a dozen of young researchers works in SMC and they can be involved when discussing publicly the results obtained in this project.

Every year in the SPD roughly 300 new students enter the undergraduate program and 200 the master course. Moreover the SPD has three Ph.D. programs (Physics, Astrophysics and Materials Science) and coordinates three more scientific Ph.D. programs in a large Ph.D. school. This is one more opportunity for the present project, which could easily attract some student interested in doing a Ph.D. thesis on the research lines described in this project.

The SPD agrees in supporting this project by offering adequate structures, e.g. the working facilities for the post-docs and the space for allocating the computer farm we plan to buy within the present project.

*** THE ROLE OF THE RESEARCH UNIT ***

Being the entire project based on a single research unit, the work to be done by this unit perfectly coincides with that described in the section "description of the research". We think is not useful to copy it again here.

We just sketch briefly the division of tasks within the unit, i.e. between the members of the research unit.

Federico Ricci-Tersenghi, given its role of scientific coordinator, will participate to all sub-projects. Nonetheless he will dedicate most of his working time to the parts of the project requiring more numerical-computational effort (SP2 and SP3).

Andrea De Martino, thanks to his experience in the field of computational biology, will be mostly responsible for SP4.

Tommaso Rizzo will be dedicated specially to analytical parts of the project and thus will contribute most to SP1.

Regarding the two post-doc researchers we plan to hire: the first one will be working in the project during the first 3 years and will be mostly dedicated to the development of algorithms, thus working in SP1 and SP2; the second one will join the project at the beginning of the second year and will be mostly concerned with application of the algorithms to problems extracted from public libraries and of biological origin, thus contributing mainly to SP3 and SP4.

All the participants will give their contribution to SP5.

We want to stress that, given the small size of the research unit, all the members will have intense exchanges and a weekly group meeting will be organized to share information and discuss together the problems found.

11 - Obiettivi scientifici del progetto di ricerca e risultati attesi

Italiano
Lo scopo principale di questo progetto è quello di costruire nuovi veloci algoritmi per la risoluzione dei problemi di inferenza statistica e ottimizzazione per modelli definiti su grafi con topologia non semplicemente aleatoria.
In altre parole, vogliamo estendere i sorprendenti risultati che sono stati ottenuti recentemente dall'applicazione di alcuni metodi della fisica dei sistemi disordinati ai problemi di inferenza e ottimizzazione su grafi aleatori: per questi problemi è stato possibile fare un incredibile salto di qualità, passando dalla risoluzione di problemi con poche migliaia di variabili a quella di problemi con alcuni milioni di variabili in tempi di calcolo confrontabili.

Riteniamo che i metodi sviluppati per trattare i modelli definiti su grafi aleatori hanno le potenzialità per essere applicati a categorie di modelli molto piú ampie. Siamo in particolare interessati a
1) reticoli regolari, su cui sono definiti la maggior parte dei modelli di fisica;
2) reti complesse, ossia con forti eterogeneità e correlazioni nelle proprie componenti.
Riguardo la seconda categoria siamo specialmente interessati alle reti che sono generate in problemi di origine biologica, per i quali esistono esperimenti che possono confermare o confutare le predizioni teoriche.

Come spiegato in modo piú dettagliato nella sezione "Descrizione della Ricerca", il presente progetto si divide in 5 sotto-progetti (SP). Riassumiamo prima di tutto in modo molto schematico quali siano gli obiettivi e i risultati attesi per ognuno di questi sotto-progetti (questi verranno meglio descritti nel seguito di questa sezione).
*** SP1 ***
Obiettivi: Integrazione dell'approssimazione CVM e del metodo delle repliche (media sull'ensemble del disordine) per vetri di spin in 2 e 3 dimensioni spaziali.
Risultati attesi: Buona descrizione della fase di alta temperatura; stima delle temperature critiche; descrizione approssimata (a livello 1RSB) della fase di bassa temperatura.
*** SP2 ***
Obiettivi: Calcolo delle probabilità marginali per uno specifico campione di vetro di spin su grafo non aleatorio con cicli (reticolo eventualmente deformato).
Risultati attesi: Algoritmo convergente di message-passing per il calcolo delle marginali in approssimazione RS e 1RSB.
*** SP3 ***
Obiettivi: Soluzione del problema di ottimizzazione a partire dalle marginali calcolate in SP2.

Risultati attesi: Studio dettagliato delle prestazioni delle procedure di decimazione e reinforcement; veloce algoritmo di ottimizzazione per le applicazioni pratiche.
*** SP4 ***
Obiettivi: Applicazione degli algoritmi sviluppati a problemi di origine biologica.
Risultati attesi: Miglioramento delle predizioni teoriche per problemi relativi alle reti metaboliche e al clustering dei dati.
*** SP5 ***
Obiettivi: Comunicazione, anche al pubblico non esperto, dei risultati ottenuti in questo progetto.
Risultati attesi: Pagine web dedicate, con informazioni scientifiche per esperti e non; software messo a disposizione con licenza GPL o CC; web server per la soluzione automatica di un problema di biologia; partecipazione a competizioni internazionali sugli algoritmi.

Nella parte piú teorico-analitica del progetto cercheremo di sviluppare le tecniche di approssimazione, quali il cluster variation method (CVM), al fine di tenere conto del grande numero di stati che un sistema disordinato ha in una fase di bassa temperatura.
Come primo passo cercheremo di coniugare la rottura della simmetria delle repliche (che descrive la struttura complessa degli stati di un vetro di spin) all'interno del CVM definito su reticoli regolari in due e tre dimensioni spaziali.

Questo sarebbe già un primo risultato molto importante visto che a 30 anni dalla definizione del modello di Edwards-Anderson (EA) che rappresenta il prototipo dei sistemi disordinati reali, cioè definiti su reticolo regolare, non esiste ancora nessuna approssimazione analitica che includa l'effetto forte di frustrazione che viene dai tanti cicli corti presenti su un reticolo regolare.
Il modello di EA, come tutti gli altri modelli disordinati su reticolo, viene studiato solo tramite simulazioni numeriche, che richiedono tempi di computazione estremamente lunghi.
Le simulazioni numeriche dei vetri di spin, e del modello di EA in particolare, sono forse tra le simulazioni numeriche piú difficili e onerose (in termini di tempi di calcolo) nel campo della meccanica statistica.
Poter usufruire di una approssimazione analitica che fornisca delle informazioni qualitativamente corrette sulle fasi termodinamiche del modello sarebbe indubbiamente un grande passo in avanti nello studio di questi modelli.

Quando si lavora con un modello disordinato, si possono studiare le sue caratteristiche fisiche mediate sull'ensemble o si possono voler misurare quelle di un campione specifico.
Noi cercheremo di sviluppare delle tecniche per rispondere ad entrambe le domande.

*** SP1 ***

Nel primo sotto-progetto (SP1) ci interesseremo al comportamento medio sull'ensemble.
Eseguendo la media sull'ensemble con il metodo delle repliche si recupera l'invarianza per traslazioni spaziali e si deve quindi lavorare con poche distribuzioni di probabilità dei messaggi (che a loro volta identificano le probabilità marginali).
Anche nella piú semplice ipotesi di simmetria sotto scambio delle repliche (replica symmetry, RS) ci aspettiamo di dover risolvere delle equazioni integrali altamente non triviali.
In particolare nel CVM che include come cluster piú grande la placchetta di quattro spin, ci aspettiamo che la struttura del problema induca nelle equazioni integrali non lineari un'operazione di de-convoluzione; questa ci porterà a sviluppare delle nuove tecniche risolutive, visto che fino ad oggi queste equazioni integrali sono state risolte con il metodo cosiddetto 'delle popolazioni' che non permette però di eseguire una de-convoluzione.

L'approssimazione RS è certamente corretta nel limite di alta temperatura e quindi riteniamo di poter fornire in quel limite del risultati corretti. Inoltre uno studio della stabilità della soluzione RS (sempre in approssimazione CVM) dovrebbe fornirci una buona stima della temperatura critica, certamente migliore di quelle analitiche note fino ad oggi, basate solamente sull'approssimazione di Bethe-Peierls.
Qualora dovessimo ottenere che in D=2 la soluzione RS è stabile per qualsiasi temperatura, avremmo nelle nostre mani un'approssimazione analitica di elevata accuratezza (visto che in D=2 il modello di EA ha una transizione di fase solo a T=0).
In D=3 ci aspettiamo che questa approssimazione fornisca una temperatura critica non nulla e riteniamo di poter migliorare ulteriormente il suo valore numerico, includendo nel CVM anche una regione cubica fatta da otto spin.

Nei casi in cui la soluzione RS all'equazioni derivate in approssimazione CVM subisce una instabilità a temperatura non nulla (plausibilmente EA in D=3), pianifichiamo di rompere la simmetria delle repliche a livello 1RSB (one-step replica symmetry breaking) e cercare una soluzione nello spazio delle distribuzioni di distribuzioni di messaggi. Questo sarà un compito molto difficile dal punto di vista numerico, ma contiamo di saperlo portare a compimento. Infatti riteniamo che il problema piú serio da affrontare nell'approssimazione CVM sia legato alla de-convoluzione e questo sarà già stato risolto a livello RS. Il passaggio alla soluzione 1RSB richiederà piú che altro un grande sforzo computazionale che sarà possibile grazie alle risorse finanziate all'interno di questo progetto.

*** SP2 ***

In un secondo sotto-progetto (SP2) ci occuperemo, invece, di studiare e risolvere uno specifico campione di un sistema disordinato. Questo problema è molto piú diretto alle applicazioni reali, visto che nei problemi pratici ci si trova a dover risolvere uno specifico problema e raramente esiste un'ensemble sul quale fare le medie.

Per semplicità all'inizio useremo campioni estratti dal modello di EA in 2 e 3 dimensioni spaziali. Le motivazioni per questa scelta sono molteplici:
1. avendo già studiato il comportamento medio sull'ensemble, avremo delle informazioni con cui confrontarci;
2. questi modelli sono stati molto studiati con simulazioni numeriche e per dimensioni ridotte esistono anche degli algoritmi esatti di risoluzione;
3. il reticolo regolare ha una topologia con tantissimi cicli corti che rendono la vita difficile agli algoritmi di tipo message-passing, e riuscire a trattare questi casi ci assicurerebbe poi un successo quasi certo su topologie un po' piú aleatorie.

Il terzo punto è quello che forse ci costringerà a lavorare su dei grafi con topologia piú "facile".
In questo caso contiamo di studiare e risolvere modelli su grafi che sono a metà tra i grafi aleatori e i reticoli regolari: ad esempio, grafi di tipo small world o tutti quei grafi che si ottengono partendo da un reticolo regolare (in 1, 2 o 3 dimensioni spaziali) e aggiungendo o ri-direzionando (rewiring) una frazione degli archi nel grafo.

In SP2 il ruolo delle simulazioni numeriche sarà determinante, dovendo infatti risolvere un numero di equazioni proporzionale alla taglia del grafo. I problemi numerici che dovremo trattare saranno piú che altro legati alla ricerca di un punto estremale dell'energia libera: a questo riguardo pensiamo di poter fornire una soluzione basata su un apposito algoritmo di message-passing. Il raggiungimento di questo obiettivo rappresenterebbe un grande risultato, permettendoci in pratica di risolvere moltissimi problemi di inferenza statistica e rappresenterebbe la base da cui partire per iniziare ad affrontare i problemi di origine biologica a cui siamo interessati.

Ovviamente esisteranno ancora molti problemi in cui l'algoritmo di message-passing derivato in approssimazione CVM-RS non sarà in grado di dare risposte. Per questo pianifichiamo fin da ora di estendere questo algoritmo a livello di una rottura di simmetria delle repliche. Questo nuovo algoritmo, sebbene sia molto piú impegnativo dal punto di vista numerico, dovrebbe essere in grado di funzionare anche nei casi in cui siano presenti correlazioni a lungo raggio.

*** SP3 ***

Il problema di ottimizzare una funzione data è strettamente legato a quello di calcolare le probabilità marginali su un dato campione: il terzo sotto-progetto (SP3) che si occuperà di problemi di ottimizzazione sarà quindi una naturale continuazione di SP2.
In SP3 verranno studiati in modo approfondito quei metodi, quali la decimazione e il 'reinforcement', che permettono di passare dalle probabilità marginali di singola variabile ad una specifica configurazione che massimizza la funzione data. Il nostro obiettivo è quello di capire sotto quali condizioni queste procedure convergano ad una buona soluzione.
La nostra risposta a questa domanda potrà essere in parte analitica per quei modelli definiti su grafi aleatori, mentre per gli altri modelli sarà basata fondamentalmente su evidenze numeriche.

SP3 ha un secondo e piú importante obiettivo: sulla base delle informazioni raccolte nello studio delle procedure di decimazione e reinforcement vogliamo proporre

un algoritmo di ottimizzazione che funzioni con alta probabilità anche su problemi derivati da applicazioni reali.
Le prestazioni di questo algoritmo verranno misurate applicandolo ai problemi di ottimizzazione contenuti in alcune ben note librerie pubbliche disponibili su internet, quali ad esempio la SATLIB.
Alla fine di questo studio pensiamo che avremo nelle nostre mani un ottimo algoritmo per l'inferenza e l'ottimizzazione. Questo algoritmo potrebbe essere un buon candidato per la soluzione dei problemi di origine biologica che studiamo in SP4.

*** SP4 ***

Il quarto sotto-progetto (SP4) si occuperà completamente dell'applicazione degli algoritmi di inferenza e ottimizzazione sviluppati in SP2 e SP3 a problemi di origine biologica. Questo è probabilmente il settore in cui, nel prossimo futuro, un miglioramento degli algoritmi di ottimizzazione e inferenza avrà l'impatto maggiore tanto a livello della ricerca pura quanto a livello applicativo, perché le reti biologiche presentano caratteristiche strutturali (per es. la struttura dei cicli) che rendono gli algoritmi tradizionali sostanzialmente inefficaci.

I problemi su cui ci concentreremo possono essere divisi in due categorie: (a) quelli relativi alle reti metaboliche, ovvero all'insieme delle reazioni chimiche con cui una cellula degrada i nutrienti (tipicamente glucosio) per sintetizzare le molecole necessarie alla sua sopravvivenza; (b) quelli relativi alle reti di regolazione trascrizionale, ovvero al complesso delle interazioni fra proteine che fungono da "interruttori" per l'espressione dei geni.

Fra i problemi del gruppo (a) che possono essere risolti con euristica "fisica", ne introduciamo qui due. Il primo riguarda il calcolo dei cicli diretti nel metabolismo, ovvero del numero di cicli in cui una particolare reazione è coinvolta. Questo equivale a calcolare il numero di modi diversi in cui la rete permette la conversione di un metabolita in un altro, ed è un indice di robustezza: è infatti lecito aspettarsi che le reazioni meno "sostituibili" siano i punti piú deboli della rete. Questa analisi consentirebbe di individuare quindi i gruppi di mutazioni piú deleterie per il funzionamento della cellula. Il secondo problema è l'individuazione dei gruppi di metaboliti conservati, ovvero delle combinazioni di metaboliti la cui concentrazione varia singolarmente ma non collettivamente. È questo un problema di grande rilevanza tanto teorica (i gruppi conservati costituiscono degli invarianti dinamici del metabolismo) quanto applicativa (la concentrazione di un metabolita può essere alterata, per es. per effetto di un farmaco, solo alterando le concentrazioni di tutti i metaboliti dello stesso gruppo) che può essere ricondotto a un problema di integer programming sul grafo del metabolismo cellulare.

Un problema centrale nel gruppo (b) è invece quello di individuare gruppi di geni coespressi in situazioni particolari (per es. tumori). Si tratta, in sostanza, di assegnare valori interi a ciascun gene in maniera tale che geni i cui profili di espressione siano simili (in qualche senso) abbiano lo stesso indice. Paragonando le assegnazioni ottenute nei casi, per es., di una cellula sana e di una cellula tumorale si possono estrarre i gruppi di geni la cui espressione è modificata dalla patologia. È naturalmente facile ottenere un clustering dei dati. Le difficoltà nascono dal fatto che i dati biologici di partenza, tipicamente da DNA chips, sono rumorosi e i clusters risultanti tendono a essere instabili rispetto a piccole perturbazioni dei dati. Con il message passing è possibile propagare informazioni da nodo a nodo della rete definita da un particolare clustering in modo tale da estrarre i cluster di massima stabilità. Questo stability-based clustering permetterebbe l'estrazione di liste di geni per la predizione dei tumori verosimilmente robuste anche in presenza di datasets limitati.

*** SP5 ***

Il quinto ed ultimo sotto-progetto (SP5) sarà dedicato a rendere pubblici i risultati di questo progetto. In particolare pensiamo di creare delle pagine web dedicate, in cui verranno descritti gli algoritmi sviluppati, verranno messi a disposizione degli utenti sia i codici (sotto licenza GPL o CC) sia le statistiche sulle loro prestazioni. Inoltre verrà istituito un web server per la soluzione automatica di un problema biologico, con l'obiettivo di convincere anche la comunità dei biologi sulla validità dei metodi basati su idee di tipo fisico.
Allo stesso tempo questi algoritmi verranno sottoposti a competizioni internazionali di carattere scientifico, quale ad esempio la SAT-competition in cui le prestazioni dei vari algoritmi di risoluzione vengono messe a confronto.

Inglese
The main objective of this project is to design new fast algorithms for solving probabilistic inference and optimization problems in models defined on graphs having a topology which is not simply that of a random graph.
In other words, we would like to extend the amazing results which has been obtained recently by applying some tools developed in the physics of disordered systems to inference and optimization problems defined on random graphs. For these problems it has been possible to make an incredible improvement, going from the resolution of problems with few thousands of variables to that of problems with millions of variables, without increasing dramatically computing times.

We believe that the tools developed for models on random graphs can be adapted to wider categories of models and graphs. We are specially interested in
1) regular lattices, where most physics models are defined;
2) complex networks, i.e. with strong heterogeneities and topological correlations.
Regarding the second class of models, we are particularly interested to networks derived from problems of biological origin, for which experimental results exist which may eventually confirm or confute our theoretical predictions.

As explained with more details in section "description of the research", the present project is divided in 5 sub-projects (SP). We first summarize in a very schematic way which are the objectives and the expected results for each sub-project (to be better explain afterwards).
*** SP1 ***
Objectives: Integration of CVM approximation and replica method (for averaging over the disorder ensemble) to study spin glasses in 2 and 3 spatial dimensions.
Expected results: Good analytical description of the high temperature phase; estimates of critical temperatures; approximate description (at the 1RSB level) of the low temperature phase.
*** SP2 ***
Objectives: Computing marginal probabilities for a specific spin glass sample on a non-random graph with loops (lattice, eventually deformed).
Expected results: Message-passing algorithm, converging with high probability, providing marginals under RS and 1RSB approximations.
*** SP3 ***
Objectives: Solving an optimization problem, given the marginals computed in SP2.
Expected results: Detailed study of the decimation and reinforcement procedures; fast optimization algorithm for dealing with practical applications.
*** SP4 ***
Objectives: Application of message-passing algorithms to problems of biological origin.
Expected results: Improving theoretical predictions for problems related to metabolic networks and data clustering.
*** SP5 ***
Objectives: Communicating widely, even to non-expert public, the results reached in this project.
Expected results: Dedicated web pages, with scientific information for experts and non-experts; software freely distributed under GPL; web server for the automatic solution of a significant biological problem; participation to international competitions on algorithms.

In the more analytical part of the project we plan to develop the approximation techniques, like the cluster variation method (CVM), in order to consider the large number of states a disordered system has in its low temperature phase.
The first step will be to merge together the breaking of the replica symmetry (which describes the complex structure o states in a spin glass) with the CVM defined on regular lattices in two and three spatial dimensions.

This would be already a very important result, given that after 30 years from the introduction of the Edwards-Anderson (EA) model (the prototype of real disordered magnetic material defined on a regular lattice) there is still no analytic approximation including the strong frustration effect coming from the many short loops on the lattice.
The EA model, as all the remaining disordered models on a lattice, can be studied only via numerical simulations, which require huge computational times.
Numerical simulations of spin glasses, and of the EA in particular, are maybe the most difficult and CPU-consuming simulations among numerical simulations in statistical mechanics.
The discovery of an analytical approximation providing a qualitatively correct picture of the phase diagram of the model would certainly be big step forward.

Working with a disordered model, one may be interested in the physical properties averaged on the ensemble or in those of a specific sample. We will try to deal with

both.

*** SP1 ***

In the first sub-project (SP1) we will be interested in studying mean properties.
The average over the disorder ensemble will be done by using the replica method; the resulting expression will be translation invariant and thus involving only few joint distributions of messages (which identify marginal probabilities).
Even under the simplest replica symmetric (RS) hypothesis, we think we will need to solve some highly non-trivial integral equations. In particular in the CVM having the 4 spin plaquette as the largest region, we believe the structure of the lattice will imply a deconvolution operation in the equations: to solve this we will develop some new method, given that the 'population dynamics' method does not allow to solve a deconvolution.

The RS approximation is certainly correct in the high temperature limit and for this reason we believe we will be able to provide very good results in that phase. Moreover a stability analysis of the RS solution (obtained under CVM) should provide us with a confident estimate of the critical temperature, certainly better than any previous estimate based on the Bethe-Peierls approximation. For example, finding that the RS solution is stable for any temperature in D=2 would confirm our solution is of very high accuracy (in D=2 the EA model has a critical point just at zero temperature). In D=3 we think we can obtain a finite critical temperature, whose numerical value can be improved by including in the CVM also the cubic region with 8 spins.

In those cases where the CVM-RS solution gets destabilized at a finite temperature (most probably in the D=3 EA model), we plan to break the replica symmetry at the level of one step (1RSB) and look for a solution in the space of distributions of distribution of messages. This search will be a very hard numerical task, but we are confident to be able to deal with it. Indeed we believe the main conceptual problem under CVM is to solve the deconvolution appearing already at the RS level; moving then to the 1RSB level is just a computational effort, that we will be able to handle thanks to the computing resources funded under this project.

*** SP2 ***

In a second sub-project (SP2) we will study and solve specific samples of a disordered system.
This problem is much more related to real applications, where one typically has a single instance to solve and rarely an ensemble of problems exist.

For simplicity we will start using samples of the EA model in 2 and 3 spatial dimensions. There are many reasons for this choice:
1. these models are studied in SP1 (taking the average over the ensemble) and we will be able t make comparisons;
2. these models have been studied numerically since long time ago and some exact solving algorithm exist which works for small systems;
3. the topology of a regular lattice contains many short loops, making the life hard to message passing algorithms, and learn how to deal with these would ensure a good performance on more random topologies.

Maybe the problem raised in point 3 above will force us to use graph with an "easier" topology (from the view point of the solving algorithm). We plan to study and solve models on graphs which interpolate between random graphs and regular lattices: for example, small world graph or all those graphs obtained starting from a regular lattice (in 1, 2 or 3 dimensions) and by adding/rewiring a fraction of the links.

In SP2 the role of numerical simulations will be central, given that we will have to solve a number of integral equations proportional to the size of the graph.
The main problem will be the search for the extremal point of the free-energy and we believe we can provide a solution based on a message passing algorithm. Reaching this goal would be a great achievement and would actually represent the starting point for approaching the problems of biological origin, we are interested in.

Obviously there will be still many problems for which the algorithm obtained under the CVM-RS approximation is not able to provide a useful answer. For this reason we plan to extend this algorithm to the 1RSB level: although the 1RSB version will be much more computer-demanding, it should work also in models with long range correlations.

*** SP3 ***

The problem of optimizing a given function is tightly related to the one of computing marginal probabilities on a given sample: thus the third sub-project (SP3), dealing with optimization problems, will be a natural continuation of SP2.
In SP3 we will make an exhaustive study of those procedures, like the 'decimation' and the 'reinforcement', which allow to go from the single variable marginal probabilities to a specific configuration maximizing the function given. Our aim is to understand under which condition these procedures do converge to a good solution. We plan to provide an analytical answer for models defined on random graphs, while for other models we will restrict to numerical evidences.

SP3 has a second and more important aim: thanks to the information we collected while studying decimation and reinforcement procedures we plan to propose an optimizing algorithm which should work with high probability also for real applications.
The performances of this algorithm will be tested by applying it to optimization problems listed in some well known public Internet libraries, e.g. SATLIB.
At the end of SP3 we will have in our hands a very good algorithm for solving inference and optimization problems. This algorithm will be a good candidate to solve problems of biological origin to be studied in SP5.

*** SP4 ***

The fourth sub-project (SP4) will be fully dedicated to the application of the inference and optimization algorithms developed in SP2 and SP3 to problems of biological origin.
This is probably the field where the improvement of algorithms in the next future will produce the most evident advance both a the level of basic research and of real applications: indeed biological networks have topological properties which make traditional algorithms largely ineffective.

The problems we shall concentrate on may be divided in two categories. (a) Those related to metabolic networks, that is to the collection of chemical reactions by which a cell processes its nutrients (typically glucose) to form the molecules needed for its survival. (b) Those related to transcriptional regulation networks, namely to the interactions between proteins that act as "switches" for gene expression.

Among the problems in group (a) that can be solved by some physical heuristics, we shall list here just two. The first one concerns the computation of cycles in metabolic networks, that is of the number of loops in which every reaction is involved. This is equivalent to computing the number of different ways in which the network converts one metabolite into another, and is a proxy for robustness: it is indeed reasonable to think that less "replaceable" reactions are the weakest points in the network. This analysis will allow for the identification of the mutations that are more deleterious for the cell by purely structural considerations. The second problem is the calculation of the conserved pools of metabolites, that is of the combinations of metabolites such that their individual concentrations vary over time while their aggregate concentration is constant. This is a problem of great practical relevance both from a theoretical viewpoint (the conserved pools are the dynamical invariants of the system) and for applications (the concentration of a certain metabolite cannot be altered, eg by a drug, without altering the concentrations of all metabolites that form a conserved pool with the target), and it can be recast in the form of an integer programming on the graph defined by cellular metabolism.

A central problem in group (b) is that of finding the groups of genes that are co-regulated in particular situations, like eg tumors. One has, essentially, to assign integer values to each gene in such a way that genes whose expression profiles are similar (in some sense) carry the same index. Comparing the assignments in the cases, for instance, of a healthy and of a cancerous cell one can extract the groups of genes whose expression is modified by the pathology. It is very easy to obtain a data clustering. The problems arise from the fact that biological data, typically obtained from DNA chips, are particularly noisy so that the resulting clusters tend to be unstable against small perturbations of the data. By message passing it is possible to propagate information from node to node of a particular clustering so as to obtain maximum-stability clusters (or sub-clusters). This stability-based clustering would permit the creation of gene lists for the prediction of tumors that will likely be robust even in the presence of small datasets.

*** SP5 ***

The fifth and last sub-project (SP5) will be dedicated to make the results of this project of public domain. We plan to create some web pages dedicated to this purpose, where algorithms will be described, the codes will be made available (under GPL or CC) and their performances compared to other algorithms.
Moreover we will build a web server to provide an automatic solution to a specific biological problem, with the aim of convincing the biology community about the validity of physics-based methods.
At the same time we will submit these algorithms to international scientific competitions, like the SAT-competition, where the performances of different algorithms are compared on the same set of hard problems.

12 - Base di partenza scientifica nazionale o internazionale

Italiano
Il problema dell'inferenza statistica e quello dell'ottimizzazione di una data funzione sono due problemi fondamentali strettamente connessi.

Data una distribuzione di probabilità definita su N variabili

$$P(x_1, ..., x_N)$$

il problema dell'inferenza statistica è essenzialmente la determinazione delle probabilità marginali su un sottoinsieme delle variabili

$$P_E(\vec{x}_E) = \Sigma_{\vec{x}_F} P(\vec{x}_E, \vec{x}_F)$$

dove E ed F sono una partizione delle N variabili. Tipicamente si considerano le probabilità marginali di singola variabile e di coppie di variabili:

$$P_i(x_i) \equiv \Sigma_{\{x_k\}_{k \neq i}} P(x_1, \ldots, x_N)$$

$$P_{ij}(x_i, x_j) \equiv \Sigma_{\{x_k\}_{k \neq i, k \neq j}} P(x_1, \ldots, x_N)$$

Gli esempi in cui il calcolo delle probabilità marginali rappresenta la soluzione del problema sono molteplici, ci limitiamo a citarne due tra i piú importanti:

1) In meccanica statistica, data una funzione hamiltoniana, l'energia libera del modello e le funzioni di correlazione possono essere scritte in termini di probabilità marginali. In particolare le marginali $P_{ij}(x_i, x_j)$ *forniscono informazioni sulla presenza di ordine a lungo raggio.*

2) Nel caso del'inferenza bayesiana, una volta note le evidenze (ad esempio i risultati di misure sperimentali) sulle variabili $\vec{x}_E$ *si vogliono calcolare le probabilità condizionate a queste evidenze*

$$P_F(\vec{x}_F | \vec{x}_E) = \frac{P(\vec{x}_E, \vec{x}_F)}{P_E(\vec{x}_E)}$$

In questa espressione il numeratore è la distribuzione di probabilità congiunta, che è nota dalla definizione modello, mentre il denominatore è la marginale sulle evidenze, ed è tipicamente la parte piú difficile da calcolare.

Il problema dell'ottimizzazione corrisponde invece al calcolo della configurazione che massimizza (o minimizza) la distribuzione di probabilità congiunta delle N variabili

$$\vec{x}^* = \text{argmax } P(x_1, \ldots, x_N)$$

Questo problema è strettamente legato (sebbene in alcuni casi non identico) a quello del calcolo delle marginali su una distribuzione

$$\tilde{P}_\beta(\vec{x}) \propto [P(\vec{x})]^\beta$$

che si concentra, nel limite $\beta \to \infty$*, vicino ai massimi della* $P(x_1, ..., x_N)$*.*

Anche questo problema appare in svariati constesti, ad esempio si è interessati:
1) in meccanica statistica, al calcolo delle configurazioni degli stati fondamentali o il calcolo dell'energia libera come minimo di una espressione variazionale;
2) nell'inferenza statistica, al calcolo della massima verosimiglianza.

Si noti che i problemi di inferenza statistica che abbiamo definito sono comuni a svariate discipline scientifiche e giocano un ruolo importante anche in molte applicazioni pratiche (si pensi ad esempio all'ottimizzazione di un processo o di una scelta intesi come minimizzazione della funzione costo o della funzione rischio). E' dunque facile intuire che un miglioramento delle tecniche di soluzione dei problemi di inferenza e ottimizzazione ha immediate ricadute in molti settori.

Purtroppo la soluzione di un problema di inferenza, tramite il calcolo esatto delle probabilità marginali, può essere ottenuta solo in alcuni casi molto particolari. Se la topologia delle interazioni tra le variabili del problema forma un albero, tutte le marginali possono essere calcolate in un tempo lineare nel numero N delle variabili.

Altrimenti, in generale il calcolo esatto richiede un tempo che cresce esponenzialmente con N e non è fattibile alle scale tipiche di molti problemi di interesse pratico. In questo caso si deve ricorrere a tecniche di approssimazione. Tra queste l'approssimazione di Bethe è attualmente lo strumento teorico più sofisticato e le sue applicazioni sono oggetto di intenso studio da parte della comunità scientifica.

Nel corso dell'ultimo decennio si è assistito alla convergenza su questo tema di diverse discipline che vanno dalla Fisica, alla Teoria dell'Informazione, all'Ottimizzazione Combinatorica e all' Intelligenza Artificiale (IA). In particolare nei primi anni novanta il campo della Teoria dell'Informazione è stato rivoluzionato dall'invenzione dei cosiddetti Turbo-Codes per la correzione di errore e dalla riscoperta dei codici low-density-parity-check (LDPC). Da allora questi codici monopolizzano l'attenzione degli esperti del settore poiché riescono ad arrivare estremamente vicino alla soglia di Shannon pur essendo molto veloci. La comprensione che questi codici sono intimamente connessi con l'approssimazione di Bethe è stata raggiunta solo in tempi recenti attraverso due passi successivi.

Innanzitutto si è riconosciuta la connessione con un algoritmo noto come Belief-Propagation (BP). Questo algoritmo è stato introdotto circa venti anni fa per risolvere il problema di inferenza statistica definito sopra [1].

Per spiegare sotto quali condizioni si può applicare questo algoritmo occorre notare che questi problemi rientrano nella classe dei Modelli su Grafi (Graphical Models). In termini fisici ciò significa che possono essere interpretati come modelli in cui le variabili sono associate a un insieme di nodi (variables nodes) di un dato grafo e le interazioni dell' Hamiltoniano sono associate un altro insieme di nodi (function nodes), in modo tale che che ad ogni nodo di interazione corrisponde un termine nell'Hamiltoniano che dipende da tutti le variabili connesse ad esso sul grafo. Il requisito fondamentale perchè si possa applicare l'algoritmo BP è che il grafo non contenga cicli. In caso contrario l'algoritmo va applicato iterativamente e la convergenza non è assicurata; peggio ancora si sa con certezza che i risultati ottenuti saranno approssimati.

Tuttavia pochi anni dopo l'invenzione dei Turbo Codes si è capito questi veloci ed efficaci algoritmi di codifica e decodifica non sono altro che applicazioni di Belief-Propagation su grafi con cicli. La scoperta che BP su grafi con cicli può dare risultati di grande precisione pur se approssimati ha avuto uno straordinario impatto [2] ed ha motivato un rinnovato interesse per questo algoritmo nella comunità che si occupa di Intelligenza Artificiale. Una grande varietà di reti bayesiane sono state studiate empiricamente e si è confermata in molti casi la straordinaria qualità delle predizioni.

Alle comunità della Teoria dell'Informazione e dell'Intelligenza Artificiale si è aggiunta nel 2000 quella dei fisici [3]. Infatti si è infine riconosciuto che BP è equivalente all'approssimazione di Bethe. Dato un modello grafico l'approssimazione di Bethe corrisponde ad assumere che la struttura del grafo intorno ad ogni dato nodo è ad albero, cioè non contiene cicli. Attraverso questa ipotesi cruciale si possono ottenere un insieme chiuso di equazioni locali che corrispondono elle equazioni iterative di BP.

In questa prospettiva si intuisce la ragione del successo di BP per i moderni codici di correzione di errore. Infatti questi codici sono strutture artificiali i cui corrispondenti modelli grafici sono definiti su grafi aleatori. I grafi aleatori contengono sì cicli che sono però tipicamente grandi se la taglia del sistema è grande: la struttura del grafo è dunque localmente ad albero, ciò che spiega perché l'approssimazione di Bethe è spesso molto buona. Meno chiaro è invece perché BP funziona bene in molte reti bayesiane che spesso presentano piccoli cicli.

Questo risultato ha prodotto interessanti sviluppi perchè è noto che l'approssimazione di Bethe si può ottenere in due modi, uno che parte da considerazioni locali, che è anche lo spirito con cui è stato introdotto Belief-Propagation, e un altro, sconosciuto alle altre comunità, che parte da un principio variazionale connesso all'energia libera del sistema. Si è così assistito ad un rilevante flusso di conoscenze tra le varie discipline. In particolare l'esportazione di una estensione dell'approssimazione di Bethe sviluppata in Fisica, il Cluster Variational Method di Kikuchi(CVM), ha portato all'introduzione dell'algoritmo Generalized-Belief-Propagation (GBP) nell'ambito dell'IA [3,4].

Nella formulazione variazionale l'approssimazione di Bethe corrisponde infatti all'ipotesi che variabili non connesse direttamente da un arco del grafo siano scorrelate. Questo permette di scrivere un'espressione approssimata dell'energia libera la cui minimizzazione porta alle equazioni iterative di BP nel caso di Bethe o di GBP nel caso di approssimazioni che assumano correlazioni a maggiori distanze (CVM).

Uno dei principali problemi aperti è quello della convergenza di BP che tipicamente cessa di convergere in determinate regioni di parametri del modello, anche utilizzando dei termini di smorzamento (damping) che generalmente aiutano molto la convergenza al punto fisso [5].
Anche in questo ambito la connessione con l'approssimazione di Bethe si è rivelata fruttuosa: il fatto che BP e GBP corrispondono alla minimizzazione di determinate approssimazioni dell'energia libera ha portato all'introduzione di algoritmi basati su tecniche di minimizzazione [4,6] che possono essere usati laddove BP o GBP non convergano.

Esistono tuttavia importanti problemi di Ottimizzazione in cui non è sufficiente passare da BP a GBP nè supplementarli eventualmente con algoritmi di minimizzazione. In alcuni di questi casi la fisica dei sistemi disordinati ha potuto spiegare l'origine del problema e fornire la soluzione, raggiungendo probabilmente il risultato più importante dell'ultimo decennio in questo campo.

Si tratta dei cosiddetti problemi a soddisfacimento di vincoli, (Constraint Satisfaction Problems, CSP) che hanno grandissima rilevanza teorica e pratica nella Computer Science. Qualora questi problemi siano definiti su grafi aleatori BP è spesso un buon algoritmo per risolverli. Tuttavia anche se i grafi sono localmente ad albero l'efficacia dell'algoritmo è limitata ad una certa regione dei parametri al bordo della quale tipicamente non c'è convergenza. Questi sistemi sono intimamente connessi alla fisica dei sistemi disordinati e questo ha permesso di comprendere l'origine di questo fenomeno [7,8]: corrisponde infatti ad una transizione di fase da una fase paramagnetica ad una fase di vetro di spin. In virtù di questa connessione uno dei maggiori risultati della fisica dei sistemi disordinati degli ultimi anni ha trovato naturale applicazione nel campo dell'ottimizzazione combinatorica. Specificamente si tratta dell'implementazione della teoria di Rottura di Simmetria delle Repliche (RSB) di Parisi nel contesto dell'approssimazione di Bethe, cioè per modelli localmente ad albero (diluiti). L'applicazione di questa teoria a problemi di ottimizzazione ha portato all'introduzione di una generalizzazione di BP nota come Survey-Propagation (SP) [9]. Questo algoritmo fornisce eccellenti risultati in molti importanti problemi di ottimizzazione combinatorica nella regione dove BP cessa di funzionare ed ha permesso di risolvere problemi di taglia inaccessibile con i metodi precedenti.

Vediamo dunque che GBP e SP affrontano due diversi problemi che possono causare l'eventuale fallimento di BP, rispettivamente la presenza di piccoli cicli nel dato modello grafico e l'emergere di una transizione di fase. Entrambi hanno mostrato di estendere in maniera significativa il range di problemi che sono trattabili e c'è unanime consenso che ulteriori progressi si dovrebbero ottenere combinando i due approcci, obiettivo che rappresenta uno dei maggiori problemi aperti.

È giusto ricordare che negli anni recenti è stato avviato anche un tentativo di migliorare l'accuratezza dell'approssimazione di Bethe, includendo sistematicamente alcuni piccoli cicli [10,11]. Tuttavia questo approccio soffre ugualmente dei problemi di convergenza di BP e quindi questo programma potrà essere continuato solo qualora vengano trovati nuovi algoritmi per migliorare le convergenza.

Riguardo le applicazioni biologiche su cui vorremmo testare i nuovi algoritmi di message-passing, ricordiamo che in campo biologico, problemi di integer programming sono molto diffusi (si pensi agli esempi riportati sopra per le reti metaboliche) ma l'applicazione di algoritmi efficienti è in fase primordiale [12,13]. Solitamente, la loro soluzione è limitata a sistemi di piccola taglia (50 reazioni), dove l'enumerazione esaustiva delle configurazioni è possibile. È il caso, per esempio del calcolo del numero di cicli in reti metaboliche piccole. Per reti piú grandi (1000 nodi, taglia realistica) questi approcci esatti sono proibitivi. Questo chiarisce ulteriormente la rilevanza che avrebbe, su diversi piani, lo sviluppo di algoritmi in grado di risolvere problemi su reti a densità di cicli alta [14].

Diverso è il caso del clustering. Esistono molti algoritmi che partizionano i dati in gruppi a partire, solitamente, dalle correlazioni. Alcuni di questi algoritmi hanno forti affinità con problemi di minimizzazione di Hamiltoniane tipo vetri di spin e richiedono tecniche Monte-Carlo piuttosto avanzate per la ricostruzione dei clusters. Recentemente, un algoritmo basato su equazioni di tipo BP (affinity propagation, e il suo derivato il soft-constrained affinity propagation) è stato implementato con relativo successo [15]. Tuttavia in casi di interesse per l'espressione genica accade che la soluzione del problema (e di conseguenza, per es., la lista di geni predittivi di un certo tumore) tenda a dipendere dall'algoritmo utilizzato, proprio per via del rumore intrinseco nei dati. Alcuni tentativi per esplorare la stabilità delle soluzioni sono stati fatti: ad esempio, aggiungendo rumore esogeno ai dati oppure clusterizzando
sottoinsiemi di dati campionati in modo appropriato. Ma tuttora la stabilità delle soluzioni (e forse i limiti intrinseci del problema) non sono stati esplorati in maniera sistematica [16].

BIBLIOGRAFIA ESSENZIALE

[1] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", San Francisco, CA: Morgan Kaufmann (1988)
[2] B.J. Frey and D.J.C. MacKay, "A Revolution: Belief Propagation in Graphs with Cycles", in NIPS (1998)
[3] J.S. Yedidia, W.T. Freeman and Y. Weiss, "Understanding Belief Propagation and its Generalizations", 2001 MERL technical report; J.S. Yedidia, W. T. Freeman and Y. Weiss, "Generalized Belief Propagation", in Advances in NIPS 13 (2001)
[4] A. Pelizzola, "Cluster variation method in statistical pkysics and probabilistic graphical models", J.Phys. A: Math. Gen. 38, R309 (2005)
[5] M. Pretti, "A message-passing algorithm with damping", J. Stat. Mech. P11008 (2005)
[6] T. Heskes, K. Albers and B. Kappen, "Approximate Inference and Constrained Optimization", in Uncertainty in Artificial Intelligence: Proceedings of the 19th Conference (UAI-2003), San Francisco: Morgan Kaufmann, 313 (2003)
[7] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman and L. Troyansky, "Determining computational complexity from characteristic `phase transitions'", Nature 400, 133 (1999)
[8] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian and L. Zdeborova, "Gibbs States and the Set of Solutions of Random Constraint Satisfaction

Problems", PNAS 104, 10318 (2007)
[9] M. Mezard, G. Parisi and R. Zecchina, "Analytic and algorithmic solution of random satisfiability problems", Science 297, 812 (2002)
[10] M. Chertkov and V.Y. Chernyak, "Loop series for discrete statistical models on graphs", J. Stat. Mech. P06009 (2006)
[11] A. Montanari and T. Rizzo, "How to compute loop corrections to the Bethe approximation", J. Stat. Mech. P10011 (2005)
[12] I. Famili and B. Palsson, "The Convex Basis of the Left Null Space of the Stoichiometric Matrix Leads to the Definition of Metabolically Meaningful Pools", Biophys. J. 85, 16 (2003)
[13] J. Wright and A. Wagner, "Exhaustive identification of steady state cycles in large stoichiometric networks", BMC Systems Biology 2, 61 (2008)
[14] C. Martelli, A. De Martino, E. Marinari, M. Marsili and I. Perez, "Identifying essential genes in E.coli by a metabolic optimization principle", PNAS 106, 2607 (2009)
[15] B.J. Frey and D. Dueck, "Clustering by passing messages between data points", Science 315, 972 (2007)
[16] L. Ein-Dor, O. Zuk, E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer", PNAS 103, 5923 (2006)

Inglese
The problem of Statistical Inference and that of the optimization of a given function are fundamental and deeply correlated.

Given a probability distribution over N variables

$$P(x_1, \ldots, x_N)$$

the statistical inference problem is essentially the computation of the marginal probability of a subset of variables:

$$P_{\mathrm{E}}(\vec{x}_{\mathrm{E}}) = \Sigma_{\vec{x}_{\mathrm{F}}} P(\vec{x}_{\mathrm{E}}, \vec{x}_{\mathrm{F}})$$

where E and F are a partition of the N variables. Typically one considers the marginal probabilities of a single variables or a couple of variables:

$$P_i(x_i) \equiv \Sigma_{\{x_k\}_{k \neq i}} P(x_1, \ldots, x_N)$$

$$P_{ij}(x_i, x_j) \equiv \Sigma_{\{x_k\}_{k \neq i, k \neq j}} P(x_1, \ldots, x_N)$$

In many cases the computation of the marginal probabilities represents the solution of the problem, for instance:

1. In statistical mechanics, given a Hamiltonian, the free energy of the model and the correlation functions can be written in terms of marginal probabilities. In particular the marginals $P_{ij}(x_i, x_j)$ *provide information on long-rang order.*

2. In Bayesian inference, once the evidences on the variable $\vec{x}_{\mathrm{E}}$ *are known (for instance the output of some experimental measures) one wants to compute the conditional probabilities*

$$P_{\mathrm{F}}(\vec{x}_{\mathrm{F}} | \vec{x}_{\mathrm{E}}) = \frac{P(\vec{x}_{\mathrm{E}}, \vec{x}_{\mathrm{F}})}{P_{\mathrm{E}}(\vec{x}_{\mathrm{E}})}$$

In this expression the numerator is the distribution of the joint probability, which is known from the definition of the model, while the denominator is the marginal over the evidences and is typically the most difficult part to compute.

The optimization problem corresponds instead to the computation of the configuration that maximize (or minimize) the probability distribution of the N variables

$$\vec{x}^{*} = \operatorname{argmax} P(x_1, \ldots, x_N)$$

This problem is tightly related (although in same cases not equivalent) to that of the computation of the marginals on a distribution

$$\tilde{P}_{\beta}(\vec{x}) \propto [P(\vec{x})]^{\beta}$$

that is peaked, in the limit $\beta \to \infty$*, around the maxima of* $P(x_1, \ldots, x_N)$.

This problem appears in many different context too, for instance:
1) in statistical mechanics one may be interested in the ground state of a given system or to the computation of the free energy as a minimum of a variational expression;
2) In statistical inference, one may be interested in the computation of the maximum likelihood function.

It should be noted that the problems we defined are common to many different fields and play a key role in many practical applications (for instance the optimization of a decision process conceived as the minimization of the cost function or the risk function). Therefore it is easy to understand that any improvement in the available techniques have important consequences in many contexts.

Unfortunately the solution to the inference problem through the exact computation of the marginal probabilities can be obtained only in a few special cases.
If the topology of the interactions between the variables is a tree, all the marginals can be computed in a time which is linear in the number of variables.
Otherwise in general the exact computation requires a time growing exponentially with N and is not feasible at the typical scales of many problems of practical interest.
In this case one has to resort to some approximation scheme. The Bethe approximation is currently the most sophisticated theoretical tool and its applications are widely studied in the scientific community.

During the last decade we have seen the convergence on this subject of various different fields, from Physics, to Information Theory (IT), Artificial Intelligence (AI) and Combinatorial Optimization.
This dates back to the early 90's when Information Theory saw the advent of the revolutionary error-correcting Turbo-Codes and the rediscovery of the low-density-parity-check (LDPC) codes. Since then those codes are the most studied by the experts because they are able to get very close to the Shannon threshold, still being very fast.
The connection between these codes and the Bethe approximation has been obtained later in two different steps.

Firstly it was recognized that the Turbo Codes are an instance of an algorithm known as Belief Propagation (BP). This algorithm was introduces almost twenty years ago in order to solve the statistical inference problem defined above [1].

To explain under what condition the algorithm can be applied it must be noted that those problems are in the class of the so-called graphical models.
From a physical point of view this means that they can be interpreted as models where the variables are associated to a class of variables nodes and the interactions of the Hamiltonian are associated to another set of nodes (function nodes) in such a way that to any interaction node corresponds a term in the Hamiltonian that depends on all the variables connected to it on the graph.
The essential requirement for the application of BP is that the graph should not contain any loop. Otherwise the algorithm has to be applied iteratively and the convergence is not certain, furthermore its predictions will be an approximation of the true marginals.

Despite this facts it was realized that the fast and precise Turbo-Codes are nothing but an instance of BP on graphs with loops. The fact that BP can give very precise results on graphical models with loops has had an extraordinary impact [2] and prompted larger interest on this algorithm in the Artificial Intelligence community. Many different Bayesian Networks with loops where studied using BP confirming in many important case the fast convergence and the high quality of the estimates.

The IT and AI communities where joined in 2000 by the physics community. It was at last recognized [3] that BP is equivalent to the Bethe approximation of Statistical Physics. On a given graphical model the Bethe approximation corresponds to the assumption that the structure of the graph around any node is tree-like. By means of this crucial hypothesis one can derive a closed set of equations equivalent to those of BP.
In this light the reason of the success of BP on the modern error correcting codes can be understood. Indeed the corresponding graphical models of these codes are defined on random graphs. Actually random graphs contain loops that are typically large when the system size is large as a result the graphs structure is locally tree-like which explain why the Bethe approximation is good. It is less clear instead why BP works fine on many Bayesian Networks that contain small loops.

This result prompted many interesting developments because it is known that the Bethe approximation can be derived in two ways. We can either start from assumption on the local structure of the graphs (which was also the original spirit of the formulation of BP) or start from an approximated expression of the variational free energy of the model.
This has produced an interesting interchange between different fields. In particular the use of an extension of the Bethe approximation, Kikuchi`s classical Cluster Variation Method (CVM) lead to the introduction of the so-called Generalized Belief Propagation (GBP) algorithm in the AI context [3,4].

In its variational formulation the Bethe approximation corresponds to the assumption that variables that are not directly connected by an edge of the graphs are uncorrelated. This allows to write down an approximate expression of free energy whose minimization leads BP in the Bethe case or to GBP in case of approximation taking into account correlations at larger distances (CVM).

One of the main open problems is the convergence of the BP algorithm which typically fails in some regions of parameters of the model, even using damping terms that should help convergence [5].
In this context the connection with the Bethe approximation was again fruitful: The fact that BP and GBP corresponds to the minimization of appropriate free energy function has led to the introduction of different algorithms based on minimization techniques [4,6] that can be applied where BP ceases to converge although are typically slower.

Nevertheless there exist some important optimization problems where it is not sufficient to use GBP instead of BP (even supplementing it with minimization procedures). In some of these cases the Physics of Disordered Systems could explain the origin of the problem and provide a solution, reaching probably the most important result in this field of the last decade.
These are the so-called Constraint Satisfaction Problem (CSP) that are of great theoretical and practical importance in Computer Science.
When these problems are defined on random graphs BP can be used to solve them. However even if the graphs are locally tree-like the algorithm can be used only in a given region of the parameters of the model while outside this region typically there is no convergence.
This systems are intimately related to the physics of disordered systems and this allowed to understand the origin of this phenomenon [7,8]: it corresponds indeed to a phase transition from a paramagnetic phase to a spin-glass phase. By virtue of this connection one of the main achievements of the theory could be applied to the field of Combinatorial Optimization. More specifically it is the extension of Parisi's Replica Symmetry Breaking in the context of the Bethe approximation, that is for models that are locally tree-like (diluted). The application of this theory to CSP led to the introduction of a generalization of BP known as Survey-Propagation (SP) [9]. This algorithm displays excellent performances in many combinatorial Optimization problems in the region where BP ceases to work and allowed to solve problem of sizes absolutely inaccessible with previous algorithms.

We see that GBP and SP tackle two different aspects that can lead to the convergence failure of BP, respectively the presence of small loops in the given graphical model and the onset of a phase transition. Both have allowed to considerably extend the range of solvable problems and there is unanimous consensus that further progresses could be achieved combining the two approaches, which is one of the main open problems.

Another open line of research is the improvement of the Bethe approximation taking care systematically of the presence of loops in the graph [10,11]. While it is known that, provided BP converges, these approaches improve the fixed point estimates, it is not clear if they can also cure the convergence problem at least in some regime.

Concerning the biological instances to which we will apply the new message passing algorithms, we note that in systems biology integer programming problems are very common (think of the examples given above for biological networks) but the application of efficient algorithms is in a primordial phase [12,13]. Typically their solution is limited to small-size systems (about 50 nodes), where an exhaustive enumeration of configurations is feasible. Such is the case for cycles in small metabolic networks. For larger systems (1000 reactions, a realistic size) exact approaches are prohibitive. This clarifies further the relevance that the design of algorithms able of coping with problems on high-cycle-density graphs would have [14].

The case of clustering is different. Many algorithms are available to partition a data set into clusters starting, normally, from the correlations. Some of these algorithms are relatives of minimization problems in Hamiltonian systems like spin glasses and use advanced Monte Carlo methods to extract the clusters. Recently, an algorithm based on BP-like equations (affinity propagation) has been implemented with relative success [15]. However in many cases of interest and particularly for gene expression the solutions of the problem (and consequently, eg, the gene list for predicting the outcome of cancer) happen to depend on the algorithm used because of the intrinsic noise in the data. Some have tried to test the stability of the clusterings, either by adding external noise to the data or by clustering subsets of the original data set. Anyway as of now there is little systematic understanding of the stability of the solutions [16].

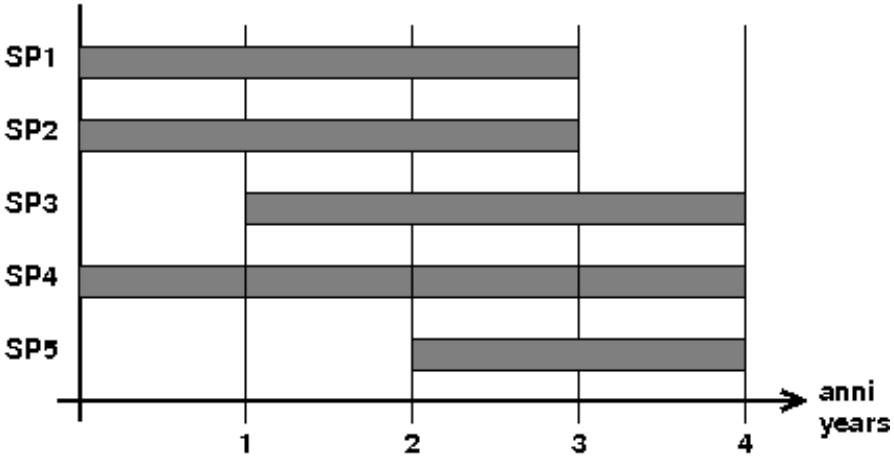
ESSENTIAL BIBLIOGRAPHY

[1] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", San Francisco, CA: Morgan Kaufmann (1988)
[2] B.J. Frey and D.J.C. MacKay, "A Revolution: Belief Propagation in Graphs with Cycles", in NIPS (1998)
[3] J.S. Yedidia, W.T. Freeman and Y. Weiss, "Understanding Belief Propagation and its Generalizations", 2001 MERL technical report; J.S. Yedidia, W. T. Freeman and Y. Weiss, "Generalized Belief Propagation", in Advances in NIPS 13 (2001)
[4] A. Pelizzola, "Cluster variation method in statistical pkysics and probabilistic graphical models", J.Phys. A: Math. Gen. 38, R309 (2005)
[5] M. Pretti, "A message-passing algorithm with damping", J. Stat. Mech. P11008 (2005)
[6] T. Heskes, K. Albers and B. Kappen, "Approximate Inference and Constrained Optimization", in Uncertainty in Artificial Intelligence: Proceedings of the 19th Conference (UAI-2003), San Francisco: Morgan Kaufmann, 313 (2003)
[7] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman and L. Troyansky, "Determining computational complexity from characteristic `phase transitions'", Nature 400, 133 (1999)
[8] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian and L. Zdeborova, "Gibbs States and the Set of Solutions of Random Constraint Satisfaction Problems", PNAS 104, 10318 (2007)
[9] M. Mezard, G. Parisi and R. Zecchina, "Analytic and algorithmic solution of random satisfiability problems", Science 297, 812 (2002)
[10] M. Chertkov and V.Y. Chernyak, "Loop series for discrete statistical models on graphs", J. Stat. Mech. P06009 (2006)
[11] A. Montanari and T. Rizzo, "How to compute loop corrections to the Bethe approximation", J. Stat. Mech. P10011 (2005)
[12] I. Famili and B. Palsson, "The Convex Basis of the Left Null Space of the Stoichiometric Matrix Leads to the Definition of Metabolically Meaningful Pools", Biophys. J. 85, 16 (2003)
[13] J. Wright and A. Wagner, "Exhaustive identification of steady state cycles in large stoichiometric networks", BMC Systems Biology 2, 61 (2008)
[14] C. Martelli, A. De Martino, E. Marinari, M. Marsili and I. Perez, "Identifying essential genes in E.coli by a metabolic optimization principle", PNAS 106, 2607 (2009)

[15] B.J. Frey and D. Dueck, "Clustering by passing messages between data points", Science 315, 972 (2007)
[16] L. Ein-Dor, O. Zuk, E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer", PNAS 103, 5923 (2006)

13 - Descrizione della Ricerca

Italiano
Il presente progetto si divide in 5 sotto-progetti (SP), che si differenziano nelle tecniche di ricerca, negli obiettivi e nella durata (illustrata nel diagramma).



*** SP1 ***

L'obiettivo principale di SP1 è quello di ottenere una integrazione tra l'approssimazione cluster variation method (CVM) e il metodo delle repliche, applicandoli ai vetri di spin definiti su reticoli regolari in 2 e 3 dimensioni spaziali.
SP1 sarà attivo per i primi 3 anni.

Volendo estendere le tecniche di inferenza e ottimizzazione (che tanto bene hanno funzionato sui grafi aleatori) a quei grafi che possiedono molti cicli corti, quali i reticoli regolari, è necessario prima di tutto partire da una approssimazione per l'energia libera che includa l'effetto almeno dei cicli piú brevi. Il cluster variation method si propone esattamente questo obiettivo e noi contiamo di usarne la versione che tiene in considerazione le regioni di spin fin alla placchetta di 4 spin: questo ci permetterà di includere esplicitamente nell'energia libera l'effetto dei cicli frustrati, tanto importanti nella fisica di un sistema disordinato.
In D=3 sarà forse necessario tentare l'inclusione anche della regione cubica di 8 spin.

L'energia libera in approssimazione CVM ancora dipende esplicitamente dal disordine 'quenched' e contiamo di usare il metodo delle repliche per eseguire la media su questo disordine. Il risultato dovrebbe essere una complicata espressione funzionale che dipende dalle distribuzioni di probabilità congiunte dei campi di cavità. A questo livello avremo ottenuto una espressione invariante sotto traslazioni spaziali e quindi dipendente solo da poche distribuzioni.

In particolare in approssimazione replico-simmetrica (RS), ossia assumendo che esista un solo stato termodinamico, ci aspettiamo l'esistenza di solo 2 distribuzioni: q(u) e Q(U,u1,u2).
Il campo di cavità u è legato alla probabilità marginale su uno spin appartenente ad una regione fatta di 2 spin, quando si esegue la somma sui valori dell'altro spin. Analogamente i campi U,u1,u2 forniscono le probabilità marginali su 2 spin appartenenti ad una regione di 4 spin, quando si somma sugli altri 2.

Passeremo quindi al calcolo dei punti estremali dell'espressione CVM-RS dell'energia libera.
Il primo tentativo sarà quello di derivare l'espressione rispetto alle distribuzioni di probabilità che vi compaiono con lo scopo di ottenere delle equazioni di punto di sella.
Purtroppo queste non saranno di facile soluzione; infatti a causa dell'approssimazione CVM compariranno nelle equazioni di punto sella delle convoluzioni del tipo

$$R(U,u1,u2) = \int dx1\, dx2\, Q(U,x1,x2)\, q(u1-x1)\, q(u2-x2)$$

che dovremo risolvere rispetto a Q(,,). Questo ci obbligherà a cercare nuovi metodi di soluzione, perché il metodo piú usato, quello delle popolazioni, non è adatto ad eseguire un'operazione di de-convoluzione.

Una strada alternativa è quella del calcolo dei punti estremali dell'energia libera. Tuttavia, a causa del metodo delle repliche, l'energia libera deve essere massimizzata in certe direzioni e minimizzata lungo altre.
Riteniamo che l'approccio migliore per risolvere questo problema sia quello di partire con lo studio della soluzione in alta temperatura: in questo regime, grazie alle simmetrie del modello, ci aspettiamo una soluzione molto piú semplice, in cui valga possibilmente u=u1=u2=0.
Lo studio delle fluttuazioni (hessiano) intorno a questa soluzione dovrebbe fornirci tutta l'informazione corretta sull'eventuale perdita di stabilità della soluzione di alta temperatura.
A questo punto potremmo avere una descrizione corretta della fase di alta temperatura e una stima molto buona delle temperature critiche del modello.

Ci muoveremo quindi nella fase di bassa temperatura, dove possiamo ottenere una prima soluzione approssimata al livello RS muovendoci sul funzionale energia libera CVM-RS nella direzione piatta indicataci dall'autovalore nullo dell'hessiano.
Una migliore approssimazione dovrebbe esserci fornita dal metodo delle repliche a livello di una rottura della simmetria delle repliche (1RSB). In questo caso il funzionale energia libera dovrebbe essere dato da un integrale su uno spazio di funzionali CVM-RS. Quella risultante sarà un'espressione molto complicata, ma riteniamo di essere in grado di fornirne alcune caratteristiche importanti (larghezza degli stati, distanza tra gli stati) almeno nelle vicinanze della temperatura critica.

*** SP2 ***

SP2 si svilupperà in perfetto parallelismo con SP1 e con questo avrà un interscambio continuo di informazioni.
In SP2 ci concentreremo su alcuni specifici campioni (scelti all'inizio in modo aleatorio e in seguito estratti da librerie pubbliche di problemi di inferenza statistica).
L'obiettivo principale di SP2 è quello di riuscire a calcolare in modo efficiente le probabilità marginali per i problemi con disordine (e frustrazione!) definiti su grafi non aleatori.

La scelta del grafo su cui lavorare sarà uno dei punti a cui dedicheremo maggior attenzione. In linea di principio vorremmo riuscire a risolvere il problema di inferenza anche per modelli definiti su reticoli regolari, ma questi hanno così tanti cicli corti che potrebbero essere "troppo" difficili per le tecniche di message-passing che vogliamo usare.

Molto probabilmente inizieremo con grafi che interpolano tra i reticoli regolari e i grafi aleatori: questo ensemble di grafi potrebbe permetterci di aumentare le correlazioni topologiche nel grafo con continuità e studiare piú sistematicamente fino a che punto i vari algoritmi di message-passing sono in grado di convergere e fornire le giuste informazioni per risolvere il problema di inferenza.

Tra gli algoritmi che pianifichiamo di usare ci sono certamente quelli già noti e basati sull'approssimazione di Bethe-Peierls: Belief Propagation (BP) e Survey Propagation (SP).
Questi algoritmi funzionano molto bene su grafi aleatori, mentre il loro comportamento su grafi con cicli corti è molto meno conosciuto (nonostante BP sia già stato usato in molti di questi grafi sotto il nome di Loopy Belief Propagation).
Contiamo di eseguire uno studio sistematico delle loro prestazioni su grafi con cicli corti.
È noto che la convergenza verso un punto fisso (soprattutto per BP) può essere migliorata con l'uso di un termine di smorzamento e ne terremo conto.

Passeremo quindi a considerare algoritmi di message-passing derivati dall'approssimazione CVM.
La prima operazione che deve eseguire un algoritmo di questo tipo è quella di capire quali sono le regioni piú importanti che vale la pena tenere in considerazione, ossia l'analogo della placchetta di 4 spin considerata principalmente in SP1. Per grafi sparsi (quali quelli che considereremo all'inizio in SP2) l'identificazione di queste regioni è molto veloce. Testeremo quindi le proprietà di convergenza a un punto fisso di questi algoritmi. In questo caso potremo studiare la probabilità di convergenza anche in funzione del numero di regioni che vengono incluse nell'approssimazione CVM.

Il naturale passo successivo è quello di tentare di includere gli effetti delle correlazioni a lungo raggio (che dovrebbero spontaneamente comparire in questi grafi a basse temperature) tramite la rottura delle repliche a livello 1RSB. In pratica questo dovrebbe produrre un algoritmo in cui i messaggi che si scambiano le regioni di CVM diventano delle distribuzioni (surveys). Questo tipo di algoritmo, sebbene sia piú oneroso dal punto di vista numerico, dovrebbe mostrare delle proprietà di convergenza migliori rispetto alla generalizzazione di BP.

La maggior parte della ricerca sviluppata all'interno di SP2 sarà strettamente numerica e per questo avremo bisogno delle risorse di calcolo previste in questo progetto. La soluzione di (decine di) migliaia di equazioni integrali non lineari è un compito estremamente esoso in termini di risorse di calcolo.

*** SP3 ***

SP3 è una naturale continuazione dei risultati ottenuti in SP2. Per questo non contiamo di iniziarlo prima della fine del primo anno del progetto e pensiamo di tenerlo attivo fino alla fine del progetto.

Il suo obiettivo principale è quello di risolvere problemi di ottimizzazione, ossia calcolare il massimo (o minimo) di una funzione F() di N variabili. Ci concentreremo su funzioni definite come somma di termini che coinvolgono un numero piccolo di variabili, ad esempio

$$F(\vec{x}) = \sum_{a=1}^{M} f_a(x_{i_{a,1}}, \ldots, x_{i_{a,k}})$$

dove M cresce proporzionalmente ad N, mentre k è un numero piccolo che non cresce con N. I termini di interazione f_a() inducono univocamente un 'factor graph' tramite il quale le variabili interagiscono e sui cui vengono scambiati i messaggi negli algoritmi di message-passing.

In linea di principio la soluzione al problema di ottimizzazione su F() è ottenibile risolvendo il problema di inferenza per la funzione

$$P(\vec{x}) \propto F(\vec{x})^{\beta}$$

nel limite $\beta \to \infty$, ossia calcolando l'energia libera nel limite di temperatura nulla.
Tuttavia in tutti quei casi in cui il massimo è degenere (problema con piú soluzioni) o quasi-degenere con altri massimi locali, le probabilità marginali non sono sufficientemente concentrate su uno dei valori che le variabili possono assumere e non ci forniscono quindi la soluzione al problema di ottimizzazione. In tutti questi casi è necessario quindi passare dalle probabilità marginali ad una singola configurazione.

Tra i metodi noti in letteratura, due ci sembrano i piú promettenti e pensiamo di studiarli entrambi:
1) la 'decimazione' assegna una variabile per volta, basandosi sui valori delle marginali, fino ad ottenere una configurazione (sperabilmente con un valore di F molto grande);
2) il 'reinforcement' introduce gradualmente un campo esterno nella direzione suggerita dalle probabilità marginali, fino a concentrare le probabilità su una unica configurazione.

Pianifichiamo di effettuare uno studio sistematico delle prestazioni di questi metodi, sugli stessi grafi usati in SP2 (ricordiamo che la convergenza degli algoritmi di message-passing applicati in SP2 è una condizione necessaria per la ricerca della configurazione ottimale).
In particolare, riteniamo di poter ottenere dei risultati importanti nella risoluzione di problemi raccolti in alcune librerie di pubblico dominio, quale ad esempio la SATLIB, migliorando sensibilmente rispetto agli algoritmi noti fino ad oggi.

*** SP4 ***

SP4 si occuperà dell'applicazione degli algoritmi studiati in SP2 e SP3 a problemi di origine biologica.
SP4 rimarrà attivo durante tutta la durata del progetto, visto che presumibilmente gli algoritmi a nostra disposizione miglioreranno durante lo sviluppo del progetto e uno stesso problema applicativo potrebbe essere affrontato e risolto in tempi successivi.

Riportiamo di seguito alcuni problemi di origine biologica che contiamo di affrontare in SP4. Premettiamo che questa lista è destinata a modificarsi, perché i partecipanti al progetto sono in contatto con alcuni importanti gruppi di ricerca che lavorano su tematiche biologiche (ad esempio il gruppo di Bialek a Princeton o quello di Vergassola all'Istituto Pasteur di Parigi) e da queste interazioni certamente scaturiranno nuove e piú interessanti linee di ricerca.

Partendo dall'analisi del numero di cicli di una rete metabolica, si comincerà dai casi piú semplici (per es. globuli rossi) di reti piccole, dove avremo la possibilità di verificare l'efficicacia degli algoritmi usati paragonandoli con quelli esatti, per poi studiare, a regime, le reti metaboliche piú complesse, come quelle batteriche (E.coli) o di eucarioti semplici (S.cerevisiae). Le predizioni sulla criticità delle reazioni verranno confrontate tanto con dati di origine clinica (ove disponibili) quanto con metodi alternativi basati sul calcolo dei flussi metabolici, che individuano nelle reazioni con variabilità minore i punti deboli della rete. Un approccio simile sarà seguito per lo studio dei gruppi di metaboliti conservati, dove tuttavia esiste la possibilità di un controllo piú diretto della validità delle predizioni effettuate. Si noti che sono disponibili i dati del metabolismo di numerosi organismi, il che consentirà analisi sistematiche.

Poco diverso è il caso del clustering. Un buon punto di partenza è dato dal contesto della affinity propagation (AP), in cui ciascun oggetto (dato) cerca fra i suoi vicini un "esempio" cui legarsi in base a una nozione di similarità predefinita. A convergenza, ciascun punto ha localizzato un esempio di riferimento e ciascun esempio corrisponde a un cluster. Si può mostrare che questo problema equivale, computazionalmente, al problema della k-median, che è NP-hard. È stato notato che questo schema, sebbene potente, soffre di scarsa robustezza: piccole perturbazioni di similarità possono portare a un riarrangiamento su larga scala dei clusters. L'introduzione di una temperatura (un costo finito per una clausola violata) migliora la situazione, al costo però di aumentare il numero di parametri. Gli algoritmi sviluppati in SP2 permetteranno di raffinare ulteriormente le soluzioni di questa classe di problemi e di introdurre, a un costo computazionale molto contenuto, un criterio di stabilità dei cluster contro piccole perturbazioni. Come sistema test useremo dati di espressione genica da DNA chips.

*** SP5 ***

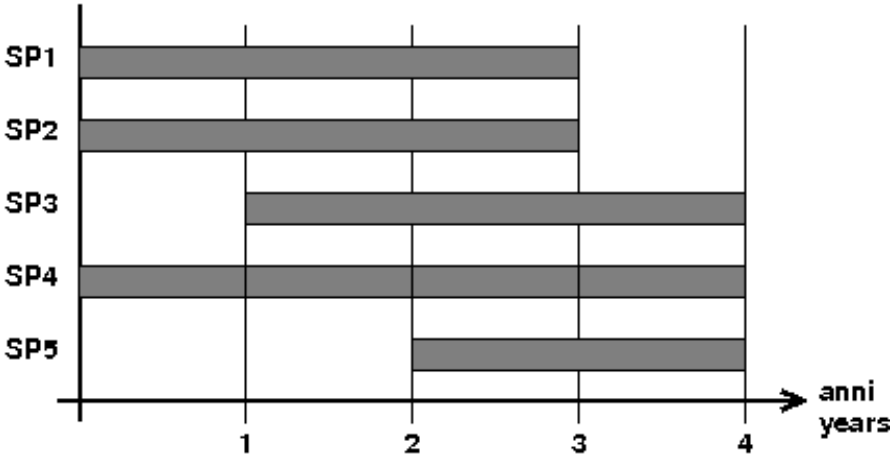
Pensiamo che dopo i primi due anni del progetto avremo nelle nostre mani un discreto numero di risultati importanti relativi alla risoluzione di problemi di inferenza e ottimizzazione, anche di interesse per le applicazioni reali.
Sarà necessario a questo punto iniziare SP5, relativo alla comunicazione dei risultati.

Con SP5 vogliamo andare oltre le normali procedure di comunicazione scientifica, tipicamente basate sullo scrivere articoli a carattere scientifico per pubblicarli su riviste specializzate. Pensiamo infatti che le problematiche trattate in questo progetto siano di interesse anche per un vasto pubblico non specializzato (si pensi ad esempio all'uso dell'inferenza statistica per fare 'screening' di malattie o all'uso dei metodi ottimizzazione per migliorare un processo produttivo).

Contiamo quindi di comunicare i risultati ottenuti all'interno di questo progetto anche verso un pubblico non specialistico. Gli strumenti principali saranno di due tipi:
1. articoli a carattere divulgativo su riviste a carattere non specialistico;
2. pagine web dedicate, in cui descrivere i problemi, gli algoritmi di soluzione e le prestazioni di tali algoritmi, confrontate con quelli noti in precedenza.
L'idea di base è quella di riuscire a convincere il pubblico interessato che un ricerca di base può spesso avere ricadute applicative apprezzabili anche da chi non è esperto del campo.

Un'altra possibile apertura del progetto verso l'esterno concerne la creazione di un web server per il calcolo dei gruppi di metaboliti conservati: l'utente fornisce in input la rete (la matrice stechiometrica) e il web server restituisce in output la struttura degli invarianti dinamici di concentrazione della rete. Questo faciliterebbe la penetrazione delle tecniche di origine fisica in ambito biologico, oltre a fornire un ovvio supporto alla ricerca. La realizzazione di questo progetto richiede alla base algoritmi che possano essere facilmente automatizzati, caratteristica spesso ovvia per gli algoritmi esatti ma meno ovvia per quelli euristici. Riteniamo che la versione di BP utile alla risoluzione del problema dei gruppi conservati possa essere, entro certi limiti, resa automatica. Questa applicazione costituisce dunque un naturale mezzo di diffusione della ricerca in ambito biofisico con un impatto potenziale piuttosto alto.

Inglese
The present project is divided in 5 sub-projects (SPs), differing in techniques, objectives and duration (see diagram).



*** SP1 ***

The principal objective of SP1 is to integrate the cluster variation method (CVM) and the replica method, applying them to spin glasses on regular lattices in 2 and 3 dimensions. SP1 will be active in the first 3 years.

If one is to generalize the inference and optimization techniques that have performed so well on random graphs to graphs with many short cycles, like regular lattices, one should first of all start from an approximation for the free energy that includes at least the effect of the shortest loops. CVM is designed for this purpose and we plan to employ a version of CVM that accounts for spin plaquettes of size up to 4: this will allow us to include explicitly in the free energy the effects due to frustrated cycles, which are crucial in the physics of a disordered system. In 3 dimensions it may be necessary to include also cubic regions formed by 8 spins.

The free energy in CVM approximation still depends explicitly on the quenched disorder and we plan to use the replica trick to calculate the average over this kind of noise. The result will likely be a complicated functional that depends on the joint probability distributions of cavity fields. At this level, we will have obtained an expression that is invariant under spatial translations and thus dependent effectively only on a few distributions.

In the replica-symmetric (RS) approximation, which assumes the existence of only one thermodynamic state, we expect there to be only two distributions: q(u) and Q(U,u1,u2). The cavity field u is linked to the marginal probability distribution on a spin belonging to a pair when the remaining spin is summed out. Similarly, the fields U,u1 and u2 give the marginal probabilities on two spins belonging to a group of four when the other two spins are summed out.

Next we will compute the extremal points of the CVM/RS expression of the free energy. Our first trial will be to derive the expression with respect to the probability distributions that appear in it, so as to obtain standard saddle point equations. Unfortunately these will be hard to solve; indeed because of the CVM approximation they will contain convolutions like

$$R(U,u1,u2) = \int dx1\, dx2\, Q(U,x1,x2)\, q(u1-x1)\, q(u2-x2)$$

which should be solved with respect to Q(,,). This will force us to look for new solution methods,since the widely used population dynamics is not suited to carry out de-convolutions.

An alternative route is the calculation of extremal points of the free energy. However, because of the replice method, the free energy should be maximized along certain directions and minimized along other directions. We believe that the best approach to this problem is that of starting from the high-temperature solution. In this regime, thanks to the model's symmetries, we expect there to be a much simpler solution, one for which u=u1=u2=0. The analysis of the fluctuations (the Hessian) around this solution should give us all the correct information about the stability of the high temperature solution. At this point we will be able to describe correctly the high temperature phase and estimate the model's critical temperature.

We will henceforth move to the low temperature phase, where it is possible to obtain a first approximate solution at the RS level by moving along the free energy functional in the direction given by the zero eigenvalue of the Hessian. An improved approximation will be given by the replica method at 1-step broken replica symmetry (1RSB) level. In this case the free energy functional is given by an integral over a space of CVM/RS functionals. The resulting expression will be very complicated but we think it is possible to characterize it to some extent (size of states, typical distance between states, etc.) at least in the vicinity of the critical temperature.

*** SP2 ***

SP2's development is parallel that of SP1. The two SPs indeed will reciprocally benefit from each other.

In SP2 we shall focus on certain specific samples (randomly chosen initially and later selected from public libraries of statistical inference problems). The main goal of SP2 is the efficient calculation of marginal probabilities for problem with disorder (and frustration!) defined on non-random graphs.

The choice of the test graph will require particular care. In principle, we would like to solve the problem of inference even for models defined on regular lattices, but such graphs have so many short loops that the kind of message passing algorithms we plan to use may be ineffective for them.

We will likely start with graphs that interpolate between regular lattices and random graphs: such an ensemble should allow for a continuous increase of the topological correlations and hence for a more systematic analysis of the limit performances of message passing algorithms for inference problems.

Among the algorithms we plan to use there are those (known) based on the Bethe-Peierls approximation: Belief Propagation (BP) and Survey propagation (SP). These algorithms work very efficiently on random graphs, while their behavior on graphs with short loops is much less clear (in spite of the fact that BP has been already used in such contexts under the name of Loopy BP). We aim at executing a systematic study of their performances on graphs with small loops. It is known that the convergence to a fixed point (especially for BP) can be improved by using a damping term and we shall count for this.

We shall then consider message passing algorithms derived from the CVM approximation. The first operation an algorithm of this type should execute is identifying the regions to be accounted for, i.e. the analogs of the 4-spin plaquette considered in SP1. For sparse graphs (those we will initially study in SP2) locating these regions is easy and fast. We shall test convergence to a fixed point in such cases. We will also be able to evaluate the convergence probability as a function of the number of regions included in the CVM approximation.

The obvious next step is including the effects of long range correlations (which should spontaneously appear at low temperatures) by breaking the replica symmetry at 1RSB level. In practice, this should produce an algorithm whose messages exchange the CVM regions becoming distributions (surveys). This kind of algorithm, while more expensive numerically, should display better convergence properties with respect to the generalization of BP.

The greatest part of the research done under SP2 will be strictly numerical and to this aim we will need the computational resources requested for this project. The solution of tens of thousands of non linear integral equations indeed involves an extremely heavy computational workload.

*** SP3 ***

SP3 is the natural continuation of SP2. For this reason we expect to activate this line of research at the end of the first year and pursue it until the project's end.

Its main objective is that of solving optimization problems, i.e. compute the maximum (or minimum) of a function F() of N variables. We will concentrate on functions defined as a sum of terms involving a small number of variables, like

$$F(\vec{x}) = \sum_{a=1}^{M} f_a(x_{i_{a,1}}, \ldots, x_{i_{a,k}})$$

where M grows linearly with N while K is a small number that does not grow with N. The interaction terms f_a() univocally induce a factor graph through which the variables interact and on which the messages in message passing algorithms are exchanged.

In principle, the solution of the optimization problem can be obtained by solving the inference problem for the function

$$P(\vec{x}) \propto F(\vec{x})^{\beta}$$

in the limit $\beta \to \infty$, i.e. computing the free energy in the limit of zero temperature. Anyhow in the cases where the maximum is degenerate (more than one solution) or almost degenerate with other local maxima, the marginal probabilities are not sufficiently peaked on one of the values that the variables can take on and hence they don't return the sought for solution. In all these cases it is necessary to move from the marginal probabilities to a single configuration.

Among the methods developed in the literature, two seem more promising to us and we plan to analyze them both:
1) decimation, which assigns one variable at a time based on marginals until a full configuration (hopefully with a very large value of F) is retrieved;
2) reinforcement, which gradually introduces an external field in the direction suggested by the marginal probabilities until the probabilities are peaked around a single configuration.

Our plan is to study the performances of these methods systematically on the same graphs used in SP2 (we remind the reader that the convergence of message passing algorithms in SP2 is a prerequisite for the search of the optimal configuration). In particular, we believe we can obtain important results in the resolution of several problems drawn from public libraries, like SATLIB, sensibly improving the current best results.

*** SP4 ***

SP4 will focus on the application of the algorithms used in SP2 and SP3 to problems of biological origin. SP4 will be active all throughout the project, given that the algorithms we will employ are likely to improve consistently as the work goes on.

We report in the following a few biological problems we aim at facing in SP4. This list will probably grow over time, also because of the existence of several collaborative links with established research groups in biological sciences (like W. Bialek's group at Princeton or M. Vergassola's group at Institut Pasteur). Such interactions naturally give rise to new and more interesting challenges.

Starting from the analysis of the number of cycles in a metabolic network, we will start from the simplest cases of small networks (e.g. red blood cells), where it will be possible to verify the effectiveness of our algorithms by comparing them to the exact ones, to address the more complicated cases like bacterial metabolic networks (e.coli) or the metabolism of simple eukarya (s.cerevisiae). The predictions on the essentiality of reactions will be tested against both clinical data (when available) and alternative methods based on the calculation of reaction fluxes, which are able to identify the critical spots in the network as the reactions with the smaller allowed variability. A similar approach will be followed for the analysis of the conserved metabolic pools, where a more direct control of the validity of our predictions is possible (through the stoichiometric matrix). Note that detailed data are available for the metabolism of many organisms, which will allow for very systematic analyses.

The case of clustering is not much different. A good starting point is given by the framework of affinity propagation (AP), where each object (data) seeks for an "exemplar" to link to among its neighbors based on a pre-defined notion of similarity. At convergence, every point has located an exemplar and each exemplar corresponds to a cluster. It can be shown that this problem is computationally equivalent to the k-median problem, which is known to be NP-hard. It was also noted that this scheme, while powerful, is not very robust as small similarity perturbations may induce large scale rearrangements of the clusters. The introduction of a temperature (a finite cost for violating a clause) improves things but only at the cost of increasing the number of parameters. The algorithms developed in SP2 will allow for a further refinement of the solutions of this class of problems and for the introduction, at a moderate computational cost, of a criterion for stability-based clustering. As a test system we will use gene expression data derived from DNA chips.

*** SP5 ***

After the first two years in the project we will have a considerable number of important results concerning the solution of inference and optimization problems, also

with direct practical relevance.

With SP5 we aim at going beyond the standard routes of scientific communication, based on writing articles for specialized scientific journals. We think that the problems we deal with in this project are of interest for a wider public (example are the use of inference methods for the screening of diseases or the use of optimization methods to improve productive processes).

We thus aim at spreading our results also to broader audiences. The main instruments will be two-fold:
1. articles on science journals for a non-specialized public
2. dedicated web pages, where problems, algorithms and their performances are explained and compared.
The basic idea is to convince the interested readers that fundamental research may have some key applicative impact that even non-experts can appreciate.

Another possible opening to the outside of this project concerns the creation of a dedicated web server for the calculation of conserved metabolic pools: the user sends the network (i.e. the stoichiometric matrix) as the input and receives as an output the structure of the dynamical concentration invariants of the network. This would greatly improve the acceptance of physical methods in the computational biological sciences, besides giving an obvious support for research. The realization of this project requires algorithms that are easily automatized, something that is often natural for exact algorithms but far less obvious for heuristic ones. We believe that the version of BP that will be used for unraveling the structure of conserved pools can be, with certain limitations, made automatic. This application is a simple and straightforward way to port our research in the biophysics community and it bears a quite high potential impact.

14 - Riassunto Spese delle Unità di Ricerca

nº	Responsabile Scientifico (codice)	Spesa A.1.1	Spesa A.1.2	Spesa A.2	Spesa B	Spesa C.1	Spesa C.2	Spesa D	Spesa E	Spesa F	Spesa G	TOTALE
1.	*RICCI TERSENGHI Federico*	*144.952*	*0*	*0*	*266.971*	*300.000*	*0*	*100.000*	*0*	*0*	*0*	811.923
	TOTALE	**144.952**	**0**	**0**	**266.971**	**300.000**	**0**	**100.000**	**0**	**0**	**0**	**811.923**

15 - Informazioni generali e durata del progetto

Durata del Progetto di Ricerca	*48 Mesi*
Mesi uomo complessivi dedicati al Progetto di Ricerca	*130*
Costo totale del Progetto	*811.923*
Finanziamento richiesto	*358.346*
Numero di contratti almeno triennali per giovani ricercatori	*2*
Costo totale	*300.000*
Numero di contratti per ricercatori di chiara fama	*0*
Costo totale	*0*

16 - Costo complessivo della Progetto di Ricerca risorse disponibili

nº	Responsabile Scientifico (codice)	Risorse finanziarie richieste al MIUR	Giovani ricercatori	Ricercatori di chiara fama internazionale	Costo totale della proposta progettuale
1.	*RICCI TERSENGHI Federico*	*358.346*	*300.000*	*0*	*811.923*
	TOTALE	**358.346**	**300.000**	**0**	**811.923**

	A carico del MIUR	A carico del Proponente	TOTALE
Costo delle attività di ricerca	*358.346*	*153.577*	511.923

Costo dei contratti almeno triennali (giovani ricercatori)	*300.000*		300.000
Costo dei contratti (ricercatori di chiara fama)	*0*		0
Costo complessivo della Progetto di Ricerca	*658.346*	*153.577*	811.923

Si ricorda che il cofinanziamento a carico del proponente deve essere pari al 30% del costo complessivo della proposta progettuale, detratti i costi dei contratti almeno triennali per giovani ricercatori e per ricercatori di chiara fama, che sono finanziati al 100%.

I dati contenuti nella domanda di finanziamento sono trattati esclusivamente per lo svolgimento delle funzioni istituzionali del MIUR. Incaricato del trattamento è il CINECA- Dipartimento Servizi per il MIUR. La consultazione è altresì riservata al MIUR - D.G. della Ricerca -- Ufficio IV, alla Commissione FIRB e ai referee scientifici. Il MIUR potrà anche procedere alla diffusione dei principali dati economici e scientifici relativi ai progetti finanziati. Responsabile del procedimento è il dirigente dell'ufficio IV della D.G. della Ricerca del MIUR.

Certifico, sotto la mia personale responsabilità, di aver ottenuto regolare autorizzazione dal rappresentante legale dell'ente di mia appartenenza, nonchè degli enti di tutte le altre Unità di Ricerca.

Firma del Coordinatore ..... Data 27/02/2009 16:39