Discrimination of Single Amino Acid Mutations of the p53 Protein by Means of Deterministic Singularities of Recurrence Quantification Analysis

Alessandro Porrello,1*, Silvia Soddu,1 Joseph P. Zbilut,2 Marco Crescenzi,3 and Alessandro Giuliani3
1Department of Experimental Oncology, Regina Elena Cancer Institute, Via delle Messi d’Oro, Rome, Italy
2Department of Molecular Biophysics and Physiology, Rush University, Chicago, Illinois
3Comparative Toxicology and Ecotoxicology Laboratory, Istituto Superiore di Sanità, Viale Regina Elena, Rome, Italy

ABSTRACT p53 is mutated in roughly 50% of all human tumors, predominantly in the DNA-binding domain codons. Structural, biochemical, and functional studies have reported that the different p53 mutants possess a broad range of behaviors that include the elimination of the tumor-suppression function of wild-type protein, the acquisition of dominant-negative function over the wild-type form, and the establishment of gain-of-function activities. The contribution of each of these types of mutations to tumor progression, grade of malignancy, and response to anticancer treatments has been so far analyzed only for a few “hot-spots.” In an attempt to identify new approaches to systematically characterize the complete spectrum of p53 mutations, we applied recurrence quantification analysis (RQA), a non-linear signal analysis technique, to p53 primary structure. Moving from the study of the p53 hydrophobicity pattern, which revealed important similarities with the singular deterministic structuring of prions, we could statistically discriminate, on a pure amino acid sequence basis, between experimentally characterized DNA-contact defective and conformational p53 mutants with a very high percentage of success. This result indicates that RQA is a mathematical tool particularly advantageous for the development of a database of p53 mutations that integrates epidemiological data with structural and functional categorizations. Proteins 2004;55:743–755. © 2004 Wiley-Liss, Inc.

Key words: database; discriminant analysis; hydrophobicity; mutation; polymorphism; prion; recurrence quantification analysis (RQA)

INTRODUCTION

The product of the p53 tumor-suppressor gene (TP53) is a transcription factor, functioning as a homotetramer, largely involved in the regulation of cell response to multiple stressing conditions and in the prevention of cancer development.1–3 Its tumor-suppressing activities are lost in the large majority of human tumors. In a number of cases this inactivation is due to alterations of different mediators or regulators of p53 pathways (e.g., MDM2 overexpression, p19ARF inactivation, p53 cytoplasmic retention).4–6 Alternatively, in about 50% of all human cancers, the cause can be a direct mutation of the TP53 gene.7,8 Remarkably, these mutations are missense in approximately 90% of cases.9 In particular, amino acid substitutions have been observed for almost two-thirds of the 393 codons of the p53 protein, including all the residues of the DNA-binding domain (DBD) (amino acids 102–292), except codons 103 and 107, which are the targets of other kinds of mutations.10,11 In contrast with this large variety of tumor mutations is the under-represented number of functionally neutral polymorphisms at the protein level in the normal population.10,11 Indeed, only one polymorphism at codon 4712 and one at codon 7213,14 whose nature is very dubious due to evidence of contribution to gain-of-function activities (see below)15 and to cancer chemotherapy modulation,16 alter the amino acid sequence of the p53 protein. Altogether, these observations point to a very peculiar p53 organizational code, characterized by a very low level of redundancy.

At a clinical-epidemiological level, the presence of p53 mutations usually results in a prognosis and therapeutic response worse than in tumors expressing the wild-type p53 protein.17–21 It has been reported that certain p53 mutations not only inactivate the wild-type tumor-suppressive functions, but also confer new protein activities. These new activities include the dominant negative function over the persisting product of the wild-type allele,22 and the development of new, oncogenic functions, the so-called gain-of-functions.23–26 A quite common criterion of classification is dependent on the subdivision of the p53 mutants affecting the DNA binding affinity into two classes. The first group is composed of mutants failing to bind the DNA due to the involvement of residues critical for DNA-protein contacts (e.g., at Arg248 and Arg273). The

---

© 2004 WILEY-LISS, INC.
second group is formed by mutants undergoing an impairment of the stable structure of the DBD (e.g., at Arg175 and His179). These two classes were defined based on crystal structure27 and biochemical studies.28–34 Interestingly, in vitro cell resistance to anticancer drugs was found to correlate with the presence of the conformational but not of the DNA-contact defective mutants in the case of etoposide, and with mutants of both groups in the case of cisplatin.25 Therefore, this group attribution is particularly relevant, since it can be linked to the cellular behavior of mutants, at least in specific cases, and has been proposed as a basis for distinct clinical approaches.35 Overall, the complex behaviors of the numerous p53 mutants suggest that a systematic analysis should be undertaken to specify their role in the etiology, pathogenesis, prognosis, and therapeutic response with respect to cancer. Nevertheless, an extensive evaluation of the structural and functional characteristics of all the mutations found in human cancers cannot be actually proposed for the very expensive and demanding work required. To overcome this problem, we asked whether it was possible to categorize the p53 mutations by means of mathematical models based on pure sequence information. From an information theory standpoint proteins can be viewed as “slightly edited strings”,36 with their sequential amino acid arrangements departing not much from randomness. Randomness, in this case, is intended in terms of substantial equivalence between the native protein sequences and their scrambled counterparts whose amino acid ordering is specified by a random-number generator algorithm. Remarkably, the only chemico-physical property showing a statistically significant non-random order is hydrophobicity. This point has been determined by means of mathematical tools such as the measure of mean-square fluctuation of hydrophobicity along the chain and of the distribution of total hydrophobicity.37 The uniqueness of hydrophobicity is probably due to the crucial role played by the peculiar patterns of juxtaposition of hydrophobic/hydrophilic residues in protein folding.38 Moreover, a considerable body of research has evidenced the importance played in protein structures of repeating amino acid motifs,39–41 as well as of complex sequence periodicities.42 Thus, we hypothesized that the unusual “fragility” of the p53 phenotype to mutations comes from a much more regular—thus more prone to be altered—pattern of hydrophobicity distribution with respect to other proteins.

In order to quantitatively assess the amount of the above mentioned regularity, use was made of recurrence quantification analysis (RQA), a computational technique that has given important results in the characterization of proteins.43–45 Here we show that the p53 hydrophobicity pattern is very similar to the prion protein (PrP) in terms of marked departure from randomness, at odds with the majority of proteins but consistently with the low p53 polymorphism. Moving from these findings, we asked whether it was possible to approach the problem of defining the characteristics of tumor-associated p53 mutants by means of RQA. This technique was able to statistically discriminate, with a very high percentage of success, the previously mentioned p53 mutations of the residues directly involved in the DNA-protein contacts and those affecting the correct folding of the p53 DBD.27 This result suggests that RQA is an appropriate tool for the development of a comprehensive p53 database, characterized by information about the structure and function of listed mutations.

THEORY AND COMPUTATION

The Model

RQA is a relatively new non-linear signal analysis technique, especially suited for short, non-stationary numerical sequences,46–48 successfully applied to the sequence of the amino acid hydrophobicities of proteins.43–45 (Hydrophobicity values are coded by the logarithm of the octanol-water partition coefficient).49 The numerical sequence corresponding to the protein primary structure is initially submitted to an n-dimensional embedding by the delay method.50 The embedding procedure consists in building an n-column embedding matrix (EM) out of the original numerical string, by shifting along the sequence of a fixed lag. Hence, the EM rows correspond to subsequent windows of length n along the sequence, and for the RQA applications must be considered like vectors of an n-dimensional space. The number of columns of the EM corresponds to the embedding dimension (emb) employed. For example, given the sequence 10, 11, 21, 32, 41, 35, 40, 19, and choosing a shifting-lag (lag) = 1 (an integer number compatible with its discrete character), the corresponding four-dimensional (i.e., emb = 4) EM is:

\[
\begin{array}{llll}
10 & 11 & 21 & 32 \\
11 & 21 & 32 & 41 \\
21 & 32 & 41 & 35 \\
32 & 41 & 35 & 40 \\
41 & 35 & 40 & 19 \\
35 & 40 & 19 & . \\
40 & 19 & . & . \\
19 & . & . & .
\end{array}
\]

In particular, the choice of using emb = 4 for general RQA applied to proteins is guided by our previous work on this field45 and by the calculation of 4 as the minimum value to extract the maximum information content encoded in the entire protein sequences.51 From the EM, RQA works out the distance matrix (DM), which reports the Euclidean distances between all the above defined vectors.46 In particular, every element \(X_{i,j}\) of the DM, which is a square matrix, corresponds to the Euclidean distance between the \(i\)-th and \(j\)-th EM rows/vectors. At this processing stage, it is highlighted the maximum Euclidean distance (maxdist) among all the calculated distances. Finally, the DM is transformed into a recurrence plot (RP), constituted only by black and white points, located at specific \(X_{i,j}\) coordinates.52 RP pixels are darkened when the corresponding distance values of DM are lower than a pre-defined radius (rad) and remain white otherwise. The features of the distance function make both the DM and the RP symmetric \((X_{i,j} = X_{j,i})\) with the latter showing a darkened main diagonal, corresponding to the
DET scaling is the graphical representation of the DET of the DET of a sequence as a function of every possible rad, going from its minimum (= 0) to its maximum (= maxdist) level (see Fig. 4). The two-dimensional DET scaling usually consists of a curve in the DET/rad plane, which converges asymptotically and relatively smoothly to 100%, owing to the increase of recurrent points proportional to rad. In contrast, for metastable proteins and peptides, it was shown that, at low rad, there is an unexpected clear divergence, with a DET spike, followed by a sudden fall and the subsequent beginning of the expected scaling. This phenomenon is called a singularity, while the interval of maxdist values in which this behavior is observed is the singularity range.

The three-dimensional DET scaling assumes a window of predefined number of consecutive amino acids (in our case 36), which we call an epoch, and repeats the two-dimensional DET scaling for all the possible epochs. It is synthetically represented by a three-dimensional graph having as axes rad, amino acid, and DET (see Fig. 5), and allows for a localization of the sequence portions of amino acids more probably involved in generating the two-dimensional DET scaling peak. Additionally, the subtraction of a three-dimensional graph from another permits to estimate how big is the mutual difference between their deterministic profiles [see Fig. 6 (a, b)].

**Local (“Sliding Window”) RQA Protocol for Mutant Discrimination**

An epoch of 36 amino acids moves along the hydrophobicity sequence of the wild type protein with lag = 1. For each of these epochs, the four RQA variables are computed and their graphs are drawn, where on the X-axis are all the epoch positions and on the Y-axis the corresponding RQA variables [see Fig. 7 (A)]. The same procedure is repeated for each mutant protein [see Fig. 7 (B)]. For each mutation the difference between homologous (i.e., representing the same RQA variable) graphs is computed [= (wild type protein) − (mutant protein)] [see Fig. 7 (C)], as well as the point by point sums of the 36 discrete values composing each difference graph. Since these graphs depend on the settled threshold (rad), this procedure is repeatedly performed in the full range 0–50% maxdist. Finally, a matrix is generated with as many rows as mutants and with four columns, corresponding to the four selected RQA variables (REC, DET, ENT, and TREND). These are tabulated as the integrals of the functions associated to the previously calculated values, in the interval 0–50% maxdist.

All the matrix values are submitted to statistical analysis, as described below.

**Statistical Analysis**

The first statistical analysis to be performed was the computation of the relative “distance from randomness” of examined proteins. This point was approached by generating a collection of twenty different scrambled versions for each protein and comparing basic RQA variables (REC and DET) of both native and scrambled sequences. The dependence of REC and DET on the amino acid order along
the chain was assessed by fixing, as a criterion of significance, a threshold of 3 standard deviations (SDs) of the difference between (D) of REC [REC D = (REC of native protein) − (average value of REC of scrambled proteins)] and DET [DET D = (DET of native protein) − (average value of DET of scrambled proteins)].

The second statistical analysis was directed toward the generation of a quantitative model discriminating p53 mutations acting by i) directly modifying residues involved in DNA-protein contacts and ii) affecting DNA binding by a general rearrangement of DBD structure. Consequently, a discriminant analysis\(^{61}\) was performed, based on the first principal component\(^{61}\) of the four selected RQA variables (REC, DET, ENT, TREND)\(^{60}\) [see Table II (a), columns 2–5]. Discriminant analysis is a statistical procedure commonly used to classify cases into the values of a categorical dependent variable by means of multiple regression. For all the mutations, the normalized difference between mutant and wild type protein was obtained to generate principal component scores [see Table II (a), column 6]. Then, for each mutant a group prediction resulted [see Table II (a), column 8] based on a discriminant score corresponding to the Mahalanobis distance from the “center of gravity” of the complete set of mutants [see Table II (a), column 9].\(^{61}\) Finally, Wilks’s lambda,\(^{61}\) a statistical test of multivariate Analysis of Variance, was applied to test the significance of the obtained discrimination between the two groups [see Table II (b)].

The statistical analysis was performed by means of StatView (SAS Institute, Inc., Cary, North Carolina).

General Set-Up, Graphical Conventions and Software for RQA

Values for RQA variables were as follows: Figures 1, 2, 3: emb = 4, line = 3, and rad = 3. The same values of emb and line were used in Figure 4, where rad varied as a percentage of maxdist, in the range 0-100. Figures 5, 6, and 7: emb = 3 and line = 2, with rad = 0–10% maxdist (Figs. 5 and 6) and = 0–50% maxdist (Fig. 7), respectively.

The values chosen are dictated by the general analysis of 1,141 proteins randomly selected from a SwissProt repository (http://www.expasy.org/sprot/), which showed an abundance of point aggregations (i.e., deterministic islands), which was over the mean levels of a standard set of proteins (our unpublished results). In particular, the largest deterministic island roughly corresponded to the amino acid positions of the PRD. To evaluate whether the minor deterministic islands are in relation with the PRD, we used the DET cross-correlation algorithm, by assuming the PRD as a reference-island. This computation, obtained by shifting along the chain a representative PRD subsequence (amino acids 68–79), showed that six out of the about twenty main deterministic islands have a pattern similar to the p53 PRD (Fig. 2).

Notably, the distribution of the deterministic islands of p53 has a high probability to be altered by mutations, compared to more common protein patterns with fewer, smaller, and less interrelated islands. In fact, whereas periodic structures of the RPs have many chances to be disorganized by amino acid substitutions, their aperiodic structures tend to remain aperiodic after analogous changes. Therefore, from a theoretical perspective, the point distribution of the p53 RP is associated with a marked susceptibility to mutations and consequently matches the p53 lack of polymorphisms.

Selection of p53 Mutants From p53 Database

A full and up-to-date p53 mutation list (final version of the database used: R8) was downloaded from the International Agency for Research on Cancer, Lyon, France (http://www-p53.iarc.fr/p53DataBase.htm). The mutations of the p53 DBD to be processed by RQA were selected on the basis of their frequency in tumors\(^{10,11}\) and of their proper characterization, as assessed by structural\(^{27,28}\) and biochemical\(^{29,30}\) studies.

RESULTS

The Hydrophobicity Pattern of p53 Reveals a Strong Deterministic Structure That Matches the p53 Lack of Polymorphisms

An analysis of the hundreds of independent mutations of p53\(^{10,11}\) has evidenced the virtual lack of any physiologically silent mutations, at the protein level, in contrast with the very high rate of tumor-associated ones. Therefore, we hypothesized that p53 possesses a self-organization characterized by a conspicuously deterministic structure, despite the apparent lack of motif regularity in its amino acid sequence, with the exception of the P’XnP repetitions of the proline rich domain (PRD) and of some amino acid couplets and triplets. To test this hypothesis, the p53 sequence of 393 amino acids was converted into a sequence of 393 hydrophobicity values and represented by a graph [Fig. 1 (a)]. This operation evidenced one region, approximately corresponding to the PRD (amino acids 58–98) [Fig. 1 (b)], where the values are included in a tight hydrophobicity interval. As expected, no further information could be obtained and no evident correlation was directly deducible among p53 regions or sub-domains by this approach. At the same time, the p53 RP [Fig. 1 (c)] showed an abundance of point aggregations (i.e., deterministic islands), which was over the mean levels of a standard set of proteins (our unpublished results). In particular, the largest deterministic island roughly corresponded to the amino acid positions of the PRD. To evaluate whether the minor deterministic islands are in relation with the PRD, we used the DET cross-correlation algorithm, by assuming the PRD as a reference-island. This computation, obtained by shifting along the chain a representative PRD subsequence (amino acids 68–79), showed that six out of the about twenty main deterministic islands have a pattern similar to the p53 PRD (Fig. 2).

Notably, the distribution of the deterministic islands of p53 has a high probability to be altered by mutations, compared to more common protein patterns with fewer, smaller, and less interrelated islands. In fact, whereas periodic structures of the RPs have many chances to be disorganized by amino acid substitutions, their aperiodic structures tend to remain aperiodic after analogous changes. Therefore, from a theoretical perspective, the point distribution of the p53 RP is associated with a marked susceptibility to mutations and consequently matches the p53 lack of polymorphisms.

p53, But Not its Family Members, Has a DET Comparable to Proteins With Diverse Kinds of Structural Complexity

Secondary and super-secondary structures have been demonstrated to have a direct correlation with the structuring of the hydrophobicity pattern.\(^{62,63}\) Nonetheless, loosely folded proteins displaying a low degree of structural complexity, exhibit markedly ordered patterns of hydrophobicity distribution.\(^{64,65}\) This point was demonstrated by an analysis of 1,141 proteins randomly selected from a SwissProt repository (http://www.expasy.org/sprot/), which showed a prominent correlation between low structural
The p53 protein is in a critical position, because it was described as intrinsically disordered with large unstructured regions, but also with a non-marginal presence of helixes and sheets in its secondary structure. In order to compare the DET of p53 (42.14) with some reference examples, we chose: the alpha chain of hemoglobin (Hb-α), the hemagglutinin neuraminidase (HN) of Sendai virus, and the PrP of the Syrian hamster. Hb-α is a globular protein with a rich and homogeneous secondary structure, and with a well-defined tertiary structuring. HN of Sendai virus is a membrane protein with a constraining hydrophobicity pattern due to its location, partially intramembrane (hydrophobic phase) and partially intracellular (hydrophilic phase). PrP is a cellular protein that can be responsible for either genetic or acquired diseases. In particular, the pathogenic protein (PrP-Sc) derives from an irreversible conformational change of the normal isoform (PrP-C). PrP was reported to belong to the group of loosely folded proteins and has a zone of high deterministic structuring that is crucial for protein–protein aggregation. The PrP sequence examined (amino acids 90–231) by RQA did not include the two regions physiologically removed in the cell (amino acids 1–22 and 232–254) and the five copper-binding octarepeat sequence. This choice took into account both the documented importance of this protein fragment and the necessity to eliminate the technical over-weighting of PrP DET, due to the presence of the octarepeats, without losing primary information on the sequence. For the sake of brevity, we will refer to this truncated form simply as PrP.

As shown in Table I, column 5, the computed value of p53 DET is comparable to these three controls. Recently, p53 was found to be member of a protein family that includes a series of different spliced forms of p63 and p73 genes. These members present sequence homologies and a certain degree of functional parallelism. Thus, we asked whether also the largest isoforms of p63 (i.e., p63x) and p73 (i.e., p73x) present a high DET like p53. Notably, the DET values give a description of p53 as a member apart (the most deterministic one) in its homology family, while the other two members have a good mutual affinity (Table I, column 5), in accordance with the experimental evidence.

The RQA Demonstrates a Strong Parallelism Between the Hydrophobicity Patterns of p53 and PrP

To clarify whether the source of high p53 DET was prevalently the presence of secondary and super-second-
ary structure (as in the case of Hb-α and HN), or of sequence motifs that create a complex DET peak (like for PrP), we tested REC and DET properties of p53 and controls. Specifically, to quantify the amount of order-dependent information conveyed by the REC and DET variables, we compared their absolute values, relatively to the original sequences, with the average REC and DET of series of twenty scrambled counterparts. Figure 3 gives a graphical appreciation of the difference between native and scrambled counterparts in the REC/DET plane, while the average values gathered from raw data are summarized in Table I. The REC D was under the significance threshold of 3 SDs for all the tabulated proteins (0 ≤ REC D ≤ 3 SDs) (Table I, columns 2, 3, 4), thus showing that their REC is scarcely sequence-related. Similar results were obtained for the DET D in the case of Hb-α, HN, p63α and p73α (0 ≤ DET D ≤ 3 SDs) (Table I, columns 5, 6, 7), as well as for a long list of other proteins (unpublished results), due to the poor structuring generally borne by protein sequences.36 In contrast, p53 and PrP have a DET D over the very selective threshold imposed (DET D > 3 SDs) (Table I, columns 5, 6, 7). Altogether, these results demonstrate that the variable DET of both p53 and PrP is markedly dependent on the amino acid sequence. Hence, the variable DET is a sensitive marker for a separation of these two proteins from other controls, based only on the deterministic disposition of their hydrophobicity patterns.

In previous works, we demonstrated that the presence of irreversible conformational switches was related to deterministic singularities in the RQA profile.59,60 These switches probably depend on the elevated flexibility of highly deterministic (repetitive) patches.39–42 PrP satisfies this deterministic requirement as well as computer-drawn peptides, designed to optimize their ability to make conformational changes.58 Remarkably, the two-dimensional DET scaling curve of p53 [Fig. 4 (a)] shows the same short-range peak singularity of PrP [Fig. 4 (b)], at odds with the other controls [Fig. 4 (c)]. Overall, these outcomes support the hypothesis that p53 and PrP deterministic patterns depend on similar structural causes,77 and are in accordance with the PrP poverty of “true” polymorphisms78–80 (see Discussion).

### Table I. REC and DET General Statistics

<table>
<thead>
<tr>
<th>Protein</th>
<th>Native REC</th>
<th>Average REC</th>
<th>REC SD</th>
<th>Native DET</th>
<th>Average DET</th>
<th>DET SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HN</td>
<td>3.58</td>
<td>3.64</td>
<td>0.22</td>
<td>40.54</td>
<td>37.75</td>
<td>2.88</td>
</tr>
<tr>
<td>Hb-α</td>
<td>4.37</td>
<td>4.60</td>
<td>0.34</td>
<td>44.79</td>
<td>40.24</td>
<td>6.73</td>
</tr>
<tr>
<td>p63α</td>
<td>2.37</td>
<td>2.05</td>
<td>0.25</td>
<td>35.93</td>
<td>29.54</td>
<td>2.52</td>
</tr>
<tr>
<td>p73α</td>
<td>3.24</td>
<td>2.85</td>
<td>0.20</td>
<td>34.64</td>
<td>33.14</td>
<td>1.46</td>
</tr>
<tr>
<td>p53 (*)</td>
<td>2.44</td>
<td>1.99</td>
<td>0.17</td>
<td>42.14</td>
<td>26.86</td>
<td>2.63</td>
</tr>
<tr>
<td>PrP (*)</td>
<td>4.37</td>
<td>3.48</td>
<td>0.34</td>
<td>47.36</td>
<td>31.84</td>
<td>3.03</td>
</tr>
</tbody>
</table>

The values of REC and DET variables of the native proteins (columns 2 and 5) are flanked by the corresponding averages (columns 3 and 6) and standard deviations (SD) (columns 4 and 7) of their scrambled counterparts. HN, hemagglutinin neuraminidase; Hb-α, alpha chain of hemoglobin; PrP, prion protein. While the difference in REC between native and scrambled sequences is under our significance threshold (≤ 3 SDs) for all the tabulated proteins, the analogous difference in DET reaches the statistical significance for both p53 and PrP (*).
of the p53 RP, thus greatly contributing in settling the properties of its deterministic pattern. The implications of this centrality are examined in the Discussion.

**Mutational “Reactivity” and Mutant Discrimination of the p53 Protein**

The above analogies between p53 and PrP suggested an examination of the systematic effects played by p53 mutations. Indeed, one of the PrP peculiarities is the ability to significantly modify the three-dimensional DET scaling even for single amino acid substitutions.59 The amount of the perturbation induced by each mutation can be graphically appreciated by subtracting the three-dimensional DET scaling graph of the mutated from that of the wild-type protein and evaluating the pseudo-volumetric change. Significantly, this procedure has evidenced a p53 situation resembling that of PrP [Fig. 6 (a, b)], because they both show important variations produced by mutations.

The holistic effects of mutations, whereby even single amino acid substitutions can interfere with the entire energy “landscape” of proteins, were investigated in studies on the structural distribution of cooperative interactions in proteins and on relationships between point mutations and protein sequence variability.81 In particular, it was shown that the broad range effect of amino acid substitutions is not due to the creation of new conformers, but rather to the setting up of a new dynamic equilibrium among the pre-existing ones. 82 Consequently, we have hypothesized that some proteins, like p53 and PrP, are more prone than others to these conformer rearrangements and that the amount of three-dimensional DET scaling perturbation is a good indicator of the protein mutational “reactivity.” Furthermore, we have postulated that RQA is intrinsically able to single out when a protein receives, by single or multiple mutations, a weak or strong impulse to the redistribution of the sub-states in the space of conformers. To test these hypotheses, we decided to study p53 mutations by using the “sliding window” RQA protocol for mutant discrimination (Fig. 7).

The p53 mutant collection subjected to our analysis consisted of 32 mutants, whose sites are located in the DBD. By crystallographic and biochemical analyses, mutants have been subdivided into two groups, failing to bind DNA owing to: i) the interference with the amino acids involved in direct contacts with DNA (group 1); ii) the
We preliminarily observed that: i) the site of the examined mutations is a unique domain; this imposition makes improbable a successful separation by RQA only due to potential differences among domains; ii) this domain is the best known and the richest in cancer-associated mutations; iii) all the sites selected have a high mutation frequency in cancer; iv) the two groups are biologically well characterized; indeed, this classification is universally recognized by the researchers in this field; v) the total number (= 32) and the group division of the examined mutations allow to perform a valid statistical study. The local RQA was able to discriminate (see Statistical Analysis of Theory and Computation) between the two groups with only two minor misclassifications (R248W and R249M) (Table II (a)). Moreover, the absolute values obtained for RQA variables allowed for the determination that global perturbations (group 2) are estimated to be...
larger than local perturbations (group 1), as results from the distribution of normalized global differences between mutant and wild type p53s [Table II (a), column 6]. Remarkably, the discriminant analysis performed (overall percentage of success: 94%) was associated with a high statistical significance (p = 0.0127), as assessed by Wilks's lambda statistics [Table II (b)]. Therefore, RQA is able to quantify the amount of perturbation given to p53 by its amino acid mutations, gathering the mutants with similar properties together. This result suggests that even particularly complex information on structure and biochemical functions of the p53 protein can be deduced by RQA, like in this case, from the mere knowledge of its amino acid sequence.

DISCUSSION

The very numerous, different mutations of the p53 protein found in a large number of human cancers have raised a series of questions concerning their biological role, clinical relevance and systematic classification. Consequently, RQA was here employed as an algorithmic tool to deal with two relevant related issues: the lack of p53 polymorphisms at the protein level and the high relevance of a global understanding of its mutants' behavior. At the same time, the strong similarity of their deterministic patterns (see Results) supported, by statistical analysis, suggests a parallelism between the p53 protein and PrP. This result strengthens a theoretical prion-tailored model of p53 describing some implications of this analogy in the set of p53 molecules present within a cell, and is in accordance with the inclusion in the same group of disordered proteins (subjected to disorder to order transitions) for p53 and PrP. Moreover, this finding links up with a work describing the capability of the p53 tetramerization domain to form amyloid-like fibrils in vitro.

The only two p53 polymorphisms observed in human beings are at amino acid 47 (P47S) and at amino acid 72 (P72R), with a population incidence low (1.5%) and variable in a wide range, respectively. Moreover, this very small number of two cases probably should be reduced to only one since it was shown that p53 mutants with a P72R polymorphism have an enhanced tumorigenic potential and a lower response rate to cancer chemotherapy. All this is in strong contrast with Chasman and Adams, who studied the polymorphisms affecting protein function and whose work allows for a demonstration that the ratio polymorphisms/total mutations is generally in the range 0.68–0.74, while for p53 it is ≤ 0.01. Even taking into account recent calls for careful evaluation of the linkage with cancer of some p53 mutations, the peculiarity of this ratio still remains. Notably, besides the closeness of their hydrophobicity patterns, the poverty of polymorphisms seems to be an additional contact point between p53 and PrP. Indeed, the same neutrality of PrP polymorphisms of the most studied animal species is extremely questionable, since it was described that polymorph PrPs are more susceptible to the transformation PrP-С → PrP-Sc. As shown (see first paragraph of the Results), the RQA interpretation of this phenomenon, relative to p53, is that it necessarily comes from the elevated structuring of its hydrophobicity pattern (highly deterministic), whose organization can be easily perturbed by amino acid substitutions, leading to pathological states. It is conjecturable that the vulnerable structure of the p53 protein is at least partially dependent on its peculiarity as a transcription factor, in which the integrity of the whole DBD, the p53 domain that holds the majority of mutations, may exert a crucial role to determine the assembly of active p53 tetramers, in addition to the functions already attributed to the p53 tetramerization domain. Interestingly, this is also consistent with the model recently sketched by Harrison et al. of the structural requirement for a β-sheet rich organization to display a prion-like behavior in terms of progressive aggregation of different protein molecules in supra-molecular assemblies. Specifically, we make the hypothesis that, as the β-structures of PrP are essential for the propagation of the pathogenic conformation, the β-structures of p53 DBD may exert a crucial role to determine the assembly of active p53 tetramers.
distance from the "center of gravity" of the full set of mutants and is normalized to 0 (corresponding to "indeterminate") with variance experimentally attributed group (7); RQA predicted group (8); score of the discriminant analysis (9). The score utilized is the Mahalanobis distance from the "center of gravity" of the full set of mutants and is normalized to 0 (corresponding to "indeterminate") with variance = 1.

### TABLE II (b). Percentages of Correct Group Attribution and Statistical Significance

<table>
<thead>
<tr>
<th>Cases attributed to group 1</th>
<th>Cases attributed to group 2</th>
<th>Cases correctly attributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA-protein contact defective mutants (group 1)</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Conformational defective mutants (group 2)</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>Total cases</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>

†Statistical significance of the classification reported in Table II (a), as measured by Analysis of Variance: Wilks's lambda statistics $F = 3.6294 (p = 0.0127)$.

Research is expected to be the study of the implications descending from the PRD cross-correlation with the sequences corresponding to the minor deterministic islands. Furthermore, the RQA finding that the high values of DET are directly related to intrinsically disordered sequences, which are important candidates to play a role in protein-protein interactions, suggests that p53 PRD may be crucial for the binding between p53 and other proteins. In particular, a relevant open question could be to verify experimentally if this domain is involved in the direct interaction of p53 with p300/CBP complexes, in analogy with what happens to SAD, the PRD of the tumor-suppressor Smad4.

### CONCLUSIONS

The biologists and physicians who are trying to define the behavior of the single p53 mutants must face relevant difficulties put by both the large mutation number and the complex experiments required in vitro and in vivo. This research points, in the long term, to the creation of a
database, which associates each mutant with its biological peculiarities, potentially helpful for the selection of different therapy protocols for oncological patients, based on the TP53 gene status. Clearly, this effort would be lessened if it were possible to cluster “a priori” the hundreds of p53 mutations recorded. Therefore, this target has become a challenging test for bio-mathematical methods such as RQA, moving from the rich public databases of p53 (http://p53.curie.fr/; http://www-p53.iarc.fr/p53DataBase.htm). In particular, our discrimination study was directed towards the DBD because this domain holds the greatest number of cancer-related mutations. The results obtained encourage applying the local RQA protocol to the entire collection of mutants, in order to analyze the grade of correlation between biological/clinical response and grade/magnitude of RQA perturbation. Regarding this proposal, the peculiar RQA profile of p53 allowed for improving the applications of our method. Indeed, the “sliding window” RQA, originally conceived as an instrument to separate allowed from non-allowed mutations of TEM-1 β-lactamases,43 was here exploited for a discrimination between two classes of mutations, obtaining a very high percentage of correct classification. As a matter of fact this is, to our knowledge, the first study on p53 that was able to perform statistically significant predictions of the biological activity of mutants based only on amino acid sequence information. It is interesting to observe that the two misclassifications (R248W and R249M) are both situated at the COOH end of the large loop three sub-domain, (one of the three large loops of this domain), whose side chains determine its whole correct spatial orientation. Therefore, we hypothesize that the prominence of these two mutations on “hot spot” codons is linked to the ability of interfering with the proper protein function even with perturbations that stay under the separation threshold of our method, at least as currently performed.

It was recently demonstrated that it is possible to counteract the action of mutations by second-site suppressor mutations carried by peptides94,95 or by direct rescue of the single site mutated proteins.96 Nonetheless, it was shown in the Results that RQA variables not only strongly correlate with energetic rearrangements, as statistically demonstrated in this work, but also that this is a positive correlation. Therefore, the heavier the structure perturbation, the bigger the modification of RQA variables. Altogether, these data support the idea that the best “normalizing” peptides or additive mutations might be planned by minimizing the difference between the RQA variables of wild type and double-mutated sequences, as assessed by our local protocol. Remarkably, the relative simplicity of this approach, potentially applicable to personalized protocols in cancer therapies, might favor a direct translation of RQA theoretical models into clinical investigation.

ACKNOWLEDGMENTS

This work was partially supported by European Union grant QLG1-1999-00273, Associazione Italiana per la Ricerca sul Cancro, Fondo per gli Investimenti della Ricerca di Base, and Ministero della Salute of Italy. This material is also based upon a joint DMS/DGMS initiative to support mathematical biology, from the National Science Foundation and National Institutes of Health, under Grant Number NSF DMS #0240230 (A.G. and J.P.Z.). A.P. is recipient of a fellowship from Fondazione Italiana per la Ricerca sul Cancro. The authors thank A. Sacchi of the Regina Elena Cancer Institute (Rome) for her kind contribution in the revision of the manuscript.

REFERENCES


ing to the infectious fragment of the scrapie isoform. Proc Natl Acad Sci USA 1997;94:10086–10091.


