Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry 29 (2005) 337–344
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

A family of metrics for biopolymers based on counting independent sets

J. Rocha*, M. Llabrés, R. Alberich

Department of Mathematics and Computer Science, Research Institute of Health Science (IUNICS), University of the Balearic Islands, 07122 Palma de Mallorca, Spain

Received 12 May 2005; received in revised form 9 August 2005; accepted 9 August 2005

Abstract

We introduce a new family of metrics for graphs of fixed size, based on counting-independent sets. Our definition is simpler and easier to calculate than the edge ideal metric family defined by Llabrés and Rosselló without loosing any of its abstract properties. We contrast them on some examples with graphs that represent protein secondary and three-dimensional (3D) structures. We conclude that although the edge ideal metrics are faster to calculate on some sparse graphs, in general, the independent set metrics are more tractable.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Counting-independent sets; Graph distance; Edge ideal metric; Hilbert polynomial; Cliques

1. Introduction

One of the main research subjects in molecular biology is to discover biopolymer functions. Biopolymers, like, for instance, proteins, are versatile molecules that are involved in the most important processes in living beings, such as catalysis, metabolism, transport, regulation, detection and destruction of invaders, etc. At the molecular level, proteins are, where life drama is (Lesk, 2001).

Since three-dimensional (3D) structure is better preserved during evolution than the primary sequence, it is possible to find proteins with less than 25% of sequence similarity but roughly the same functions and tri-dimensional shape. When this tri-dimensional structure is represented as a graph, we can compare structures of different proteins and find the principal properties that define a structure. However, it is worth noticing that occasionally, proteins with similar shapes have not known similar functionality. Among more than 30,000 proteins in the Protein Data Bank (Berman et al., 2000) only 2000 are structurally different. In addition, it is known that the spacial distribution of certain amino acids on the protein surface determines its association with other molecules, and therefore, its function. From all of this, we can deduce that the algorithmic support for protein classification receives as input mainly the 3D coordinates of the protein atoms and in this matter these algorithmic problems have no difference with other problems of computational geometry, computer vision, computer graphics, medical imaging, etc.

Our efforts are concentrated on comparing structures that allow the use of not only the primary structure of biopolymers, including RNA molecules and proteins, but also arbitrary relationships between the monomers or atoms, like chemical bonds between them or properties like closeness or visibility.

The similarity of structures can be quantified by means of metrics. In order to use such a metric, a biopolymer is modelled as a graph whose vertex set is sorted and whose arcs represent contacts (bonds) between monomers. Reidys and Stadler (1996) defined metrics on RNA secondary structures of fixed length and unique bonds (a base can only pair with at most one base) using involution and permutation subgroups.

Then, the edge ideal metrics family for biopolymers was introduced, and it is considered the first algebraic metric family for biopolymers (Llabrés and Rosselló, 2003). It represents the graph structure using edge ideals. Its main results

* Corresponding author. Tel.: +34 971172969; fax: +34 971173003.
E-mail addresses: jairo@uib.es (J. Rocha), merce.llabres@uib.es (M. Llabrés), r.alberich@uib.es (R. Alberich)

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.08.003

are the possibility to compare arbitrary graphs of fixed length without unique bonds, and the method to calculate them expressed in terms of Hilbert polynomials, which are computed using several public domain computer algebra systems like, for instance, CoCoA (Capani et al., 2005) and Macaulay (Grayson and Stillman, 2005).

However, to calculate a Hilbert polynomial of the edge ideal of a graph needs an exponential algorithm in the number of edges, unless $\mathbf{P} = \mathbf{NP}$ (Bayer and Stillman, 1992). Besides, the algorithm itself is complex (e.g., long) and available only on a specialised computer algebra package.

We propose in this paper a new family that is analogous to the edge ideal metric family but that a fix (low-order) member of the family can be calculated polynomially (efficiently) and that is simpler to define, based on the better studied problem of counting-independent sets (Jou and Chang, 2000); our approach is combinatorial rather than algebraic. The metrics are also uniformly related among themselves and able to manage structures with no unique contacts.

In this paper, we also show some examples of application on proteins. More applications of these metrics to measure the similarity of biopolymers of the same length are likely to come from attempts to compare different foldings or secondary structures of the *same* molecule (Moulton et al., 2000; Zuker, 1989; Schuster et al., 1994).

2. Metric definitions

In this section, we recall some definitions on graphs and define the new metrics.

Let $G = (V, E)$ be an undirected graph without self-loops. An *independent set* X of G is a subset of vertices that are pairwise not related by E, i.e., such that if $i, j \in X$, then, $\{i, j\} \notin E$. A *clique* K of G is a subset of vertices pairwise related by E. Notice that K is a clique of G if an only if K is an independent set of G^c, the graph with the same vertices V but with exactly all the edges that are not in E.

Let $\mathcal{I}(G)$ be the set of all independent sets of G, $\mathcal{I}(G)_m$ the set of all independent sets of G of size $\leq m$ and $\mathcal{I}(G)^{(m)}$ the set of all independent sets of G of size m. We also define, $I_G(m) = |\mathcal{I}(G)_m|$ and $\mathrm{ind}_G(m) = |\mathcal{I}(G)^{(m)}|$.

Let n be from now on an integer greater than 1. We shall denote by $\mathcal{C}_n$ the set of graphs that have $V = \{1, \ldots, n\}$ as vertex set. Given two graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$, their *union* is the graph $G_1 \cup G_2 = (V, E_1 \cup E_2)$.

Example 1. Let $n = 8$ and

$$G_1 = (V, \{\{2, 6\}, \{3, 7\}, \{4, 8\}\}),$$
$$G_2 = (V, \{\{2, 6\}, \{3, 7\}, \{4, 8\}, \{2, 8\}\})$$

be the graphs shown in Fig. 1. The first one could be a representation of an α helix of length 3 in a protein, as explained in Section 4. The second graph has one additional edge. Then,

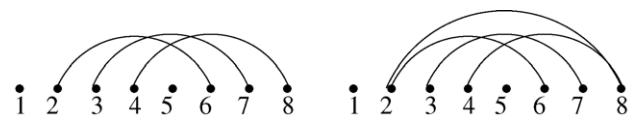

Fig. 1. Graphs G_1 and G_2 of Example 1.

$$\begin{aligned}\mathcal{I}(G_1)^{(5)} = \ & \{\{1, 2, 3, 4, 5\}, \{1, 6, 3, 4, 5\}, \{1, 2, 7, 4, 5\}, \\ & \{1, 6, 7, 4, 5\}, \{1, 2, 3, 8, 5\}, \{1, 6, 3, 8, 5\}, \\ & \{1, 2, 7, 8, 5\}, \{1, 6, 7, 8, 5\}\}, \\ \mathcal{I}(G_2)^{(5)} = \ & \{\{1, 2, 3, 4, 5\}, \{1, 6, 3, 4, 5\}, \{1, 2, 7, 4, 5\}, \\ & \{1, 6, 7, 4, 5\}, \{1, 6, 3, 8, 5\}, \{1, 6, 7, 8, 5\}\}.\end{aligned}$$

Notice that $\mathcal{I}(G_2)^{(5)}$ is the same set $\mathcal{I}(G_1)^{(5)}$ except for the two subsets that contain the extra edge $\{2, 8\}$. We have the following values:

m	0	1	2	3	4	5	6	7	8
$\mathrm{ind}_{G_1}(m)$	1	8	25	38	28	8	0	0	0
$\mathrm{ind}_{G_2}(m)$	1	8	24	34	23	6	0	0	0

Definition 1. For every $G_1, G_2 \in \mathcal{C}_n$ and for every m, $2 \leq m \leq n$,

$$i'_m(G_1, G_2) = |\mathcal{I}(G_1)_m \Delta \mathcal{I}(G_2)_m|.$$

Proposition 2. *i'_m as defined above is a metric on $\mathcal{C}_n$.*

Proof. Let us start showing that if $i'_m(G_1, G_2) = 0$, then, $G_1 = G_2$. Indeed, if there is an edge $\{j, k\}$ in G_1 but not in G_2, then, $\{j, k\}$ is an independent set of size 2 in G_2 but not in G_1 that will belong to $\mathcal{I}(G_1)_m \Delta \mathcal{I}(G_2)_m$ for $m \geq 2$, making $i'_m(G_1, G_2) \geq 1$.

The triangular inequality holds because for any sets A, B and C, $(A \Delta B) \subseteq (A \Delta C) \cup (C \Delta B)$, and then, $|A \Delta B| \leq |(A \Delta C) \cup (C \Delta B)| \leq |A \Delta C| + |C \Delta B|$.

The other metric properties are obvious. □

The following remark is useful.

Remark 1. For every $G = (V, E)$, $G' = (V, E')$,

$$i'_2(G, G') = |E \Delta E'|.$$

Indeed, $\mathcal{I}(G)^{(1)} = \mathcal{I}(G')^{(1)}$ (all vertex singletons) and $\mathcal{I}^{(2)}(G) = E^c$, $\mathcal{I}^{(2)}(G') = E'^c$ (all vertex pairs that are not edges), then, $\mathcal{I}(G)_2 \Delta \mathcal{I}(G')_2 = E^c \Delta E'^c = E \Delta E'$, as we claimed.

And the next one will be used to deduce a very useful formula.

Remark 2. For every $G_1, G_2 \in \mathcal{C}_n$ and for every $m,\ 2 \le m \le n$

$$\mathcal{I}(G_1)_m \cap \mathcal{I}(G_2)_m = \mathcal{I}(G_1 \cup G_2)_m.$$

It is enough to realize that when X is an independent set of G_1 and G_2, neither G_1 nor G_2 contain an edge relating a pair of elements of X. But this is clearly equivalent to the fact that $G_1 \cup G_2$ does not contain any edge relating a pair of elements of X, i.e., to the fact that X is an independent set of $G_1 \cup G_2$.

Proposition 3.

$$i'_m(G_1, G_2) = I_{G_1}(m) + I_{G_2}(m) - 2I_{G_1 \cup G_2}(m).$$

Proof. We have that

$$\begin{aligned}
&|\mathcal{I}(G_1)_m \Delta \mathcal{I}(G_2)_m| \\
&\quad = |\mathcal{I}(G_1)_m - (\mathcal{I}(G_1)_m \cap \mathcal{I}(G_2)_m)| + |\mathcal{I}(G_2)_m \\
&\quad - (\mathcal{I}(G_1)_m \cap \mathcal{I}(G_2)_m)| = |\mathcal{I}(G_1)_m - (\mathcal{I}(G_1 \cup G_2)_m)| \\
&\quad + \mathcal{I}(G_2)_m - (\mathcal{I}(G_1 \cup G_2)_m)| = |\mathcal{I}(G_1)_m| \\
&\quad - |\mathcal{I}(G_1 \cup G_2)_m| + |\mathcal{I}(G_2)_m| - |\mathcal{I}(G_1 \cup G_2)_m| \\
&\quad = (I_{G_1}(m) - I_{G_1 \cup G_2}(m)) + (I_{G_2}(m) - I_{G_1 \cup G_2}(m)),
\end{aligned}$$

as we claimed. □

The metrics i'_m grow with n and m, and thus, it is convenient to normalise them in order to avoid unnecessarily high figures. More specifically, let $G_0 \in \mathcal{C}_n$ be the graph on n vertices with no edges, and $G' \in \mathcal{C}_n$ be a graph with only one edge, say the edge $\{j, k\}$. Every subset X of vertices is independent of G_0, while it is not independent of G' if and only if $\{j, k\} \not\subseteq X$. Then, the subsets X that are independent sets of one graph but not of the other are the ones that contain $\{j, k\}$. To have one such an independent set of size l, we could choose $l - 2$ vertices from $n - 2$ nodes. Therefore, $|\mathcal{I}(G_0)^{(l)} \Delta \mathcal{I}(G')^{(l)}| = \binom{n-2}{l-2}$ and

$$\begin{aligned}
i'_m(G_0, G') = |\mathcal{I}(G_0)_m \Delta \mathcal{I}(G')_m| &= \sum_{l=2}^{m} |\mathcal{I}(G_0)^{(l)} \Delta \mathcal{I}(G')^{(l)}| \\
&= \sum_{l=2}^{m} \binom{n-2}{l-2}.
\end{aligned}$$

Let us define this number as

$$a_m = \sum_{l=2}^{m} \binom{n-2}{l-2}.$$

If we take 1 as the "natural" value for the distance between G_0 and G', then, instead of using the metrics i'_m, we divide them by a_m.

Definition 2. For every $m,\ 2 \le m \le n$, the independent set mth metric on $\mathcal{C}_n$ is

$$i_m(G_1, G_2) = \frac{1}{a_m} i'_m(G_1, G_2), \quad G_1, G_2 \in \mathcal{C}_n.$$

Example 4. If we continue with Example 1, since $E_1 \subset E_2$, the formula of Proposition 3 is reduced to

$$i'_m(G_1, G_2) = I_{G_1}(m) - I_{G_2}(m).$$

We have the following values:

m	$\mathrm{ind}_{G_1}(m)$	$\mathrm{ind}_{G_2}(m)$	$I_{G_1}(m)$	$I_{G_2}(m)$	$i'_m(G_1, G_2)$	a_m	$i_m(G_1, G_2)$
2	25	24	34	33	1	1	1.00
3	38	34	72	65	5	7	0.71
4	28	23	100	90	10	22	0.45
5	8	6	108	96	12	42	0.29
6	0	0	108	96	12	57	0.21
7	0	0	108	96	12	63	0.19
8	0	0	108	96	12	64	0.19

After this normalisation, these independent set metrics satisfy the following property.

Proposition 5. *Let $G_1, G_2 \in \mathcal{C}_n$. For every $m,\ 2 \le m \le n$,*

$$\frac{a_m}{a_{m+1}} i_m(G_1, G_2) \le i_{m+1}(G_1, G_2) \le i_m(G_1, G_2).$$

To prove it, we need some lemmas. We fix $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ in $\mathcal{C}_n$, and for simplicity, use the following notation:

$$\mathcal{I}^{(m)} = \mathcal{I}(G_1)^{(m)} \Delta \mathcal{I}(G_2)^{(m)},$$
$$i'_m = i'_m(G_1, G_2).$$

Lemma 6. *For every $k,\ 2 \le k \le n$,*

$$(k-1)|\mathcal{I}^{(k+1)}| \le (n-k)|\mathcal{I}^{(k)}|.$$

Proof. Let us suppose that $X \in \mathcal{I}^{(k+1)}$ is an independent set of G_1 but not of G_2. There are at least $k - 1$ different vertices u that can be removed from X so that $X - \{u\}$ is still not an independent set of G_2, and thus, $X - \{u\} \in \mathcal{I}^{(k)}$. Indeed, if $v, w \in X$ are such that $\{v, w\} \in E_2$ (they exist because X is not an independent set of G_2), then, we can remove any $u \in X - \{v, w\}$ and the resulting set is still not independent of G_2 because v and w are related. For each $X \in \mathcal{I}^{(k+1)}$, let us fix one such subset $X' = X - \{v, w\} = \{u_1, \ldots, u_{k-1}\}$ of ordered vertices such that $X - \{u_j\} \in \mathcal{I}^{(k)}$, for every $j = 1, \ldots, k-1$. We do the same if X is an independent set of G_2 but it is not of G_1.

Now, we consider the mapping $\phi : \{1, \ldots, k-1\} \times \mathcal{I}^{(k+1)} \to \{1, \ldots, n-k\} \times \mathcal{I}^{(k)}$ defined by:

$$\phi(j, X) = (l, X - \{u_j\})$$

where l is defined as follows: let $\{1, \ldots, n\} - (X - \{u_j\}) = \{w_1, \ldots, w_{n-k}\}$, the ordered vertex list of the complement of X together with u_j; then, l is the index such that $w_l = u_j$. Notice that ϕ is well defined because $X - \{u_j\} \in \mathcal{I}^{(k)}$ by construction.

To prove the inequality, let us see that ϕ is injective. Assume that $\phi(j_1, X) = \phi(j_2, Z) = (l, Y)$. Then, $Y = X - \{u_{j_1}\} = Z - \{v_{j_2}\}$, where u_{j_1} and v_{j_2} are in $X' \subseteq X$ and $Z' \subseteq Z$, the subsets fixed, respectively, for X and Z. Besides, in the ordered list $Y^c = \{w_1, \ldots, w_{n-k}\}$, $w_l = u_{j_1} = v_{j_2}$. Therefore, $X = Z$: if $v \in X$, $v \neq w_l$, then, $v \in Z$ because $X - \{w_l\} = Z - \{w_l\}$; on the other hand, if $v = w_l$, then, $v \in Z$ also, and the same argument for the other direction $Z \subseteq X$. Since $X = Z$, $X' = Z'$ and $j_1 = j_2$. This concludes the proof. □

Lemma 7. *For every m, $2 \leq m \leq n$,*

$$a_m |\mathcal{I}^{(m+1)}| \leq \binom{n-2}{m-1} i'_m$$

Proof. First, let us prove by induction on m that for every j and m, $2 \leq j \leq m \leq n$,

$$\binom{n-2}{j-2} |\mathcal{I}^{(m+1)}| \leq \binom{n-2}{m-1} |\mathcal{I}^{(j)}|.$$

If $m = 2$, $j = 2$, the inequality is reduced to the previous lemma with $k = 2$.

We will use next that

$$|\mathcal{I}^{(m+2)}| \leq \frac{n-m-1}{m} |\mathcal{I}^{(m+1)}|,$$

which is a consequence of the previous lemma with $k = m + 1$.

Assume now the inequality to be true for m and for every $j \leq m$, and let us prove it for $m + 1$ and for every $j \leq m + 1$. If $j \leq m$, then,

$$\begin{aligned}\binom{n-2}{j-2} |\mathcal{I}^{(m+2)}| &\leq \binom{n-2}{j-2} \tfrac{n-m-1}{m} |\mathcal{I}^{(m+1)}| \\ &\leq \tfrac{n-m-1}{m} \binom{n-2}{m-1} |\mathcal{I}^{(j)}|,\end{aligned}$$

by the induction hypothesis,

$$= \binom{n-2}{m} |\mathcal{I}^{(j)}|,$$

as needed. In the case $j = m + 1$, we cannot use the induction hypothesis, but it is again a consequence of the previous lemma for $k = m + 1$; this lemma, for $k = m + 1$ says

$$m|\mathcal{I}^{(m+2)}| \leq (n - m - 1)|\mathcal{I}^{(m+1)}|$$

and then, if we multiply both sides of the inequality by $\frac{(n-2)(n-3)\cdots(n-m)}{m(m-1)\cdots 2\cdot 1}$, we obtain

$$\binom{n-2}{m-1} |\mathcal{I}^{(m+2)}| \leq \binom{n-2}{m} |\mathcal{I}^{(m+1)}|$$

as desired. Now adding up both sides of the proved inequality for $j = 2 \ldots, m$, we have

$$\sum_{j=2}^{m} \binom{n-2}{j-2} |\mathcal{I}^{(m+1)}| \leq \binom{n-2}{m-1} \sum_{j=2}^{m} |\mathcal{I}^{(j)}|,$$

as we claimed. □

Now, we prove Proposition 5.

Proof. Notice that

$$i'_{m+1} = i'_m + |\mathcal{I}^{(m+1)}|.$$

Hence, we have that

$$i'_{m+1} \leq i'_m \left(1 + \frac{\binom{n-2}{m-1}}{a_m} \right),$$

using the previous lemma. The last factor

$$1 + \frac{\binom{n-2}{m-1}}{a_m}$$

is equal to $\frac{a_{m+1}}{a_m}$, exactly what we need to prove that $i_{m+1}(G_1, G_2) \leq i_m(G_1, G_2)$.

The other inequality is easier and follows from the fact that $i'_m \leq i'_{m+1}$, and the normalisation factors:

$$\begin{aligned}\frac{a_m}{a_{m+1}} i_m(G_1, G_2) = \frac{i'_m(G_1, G_2)}{a_{m+1}} &\leq \frac{i'_{m+1}(G_1, G_2)}{a_{m+1}} \\ &= i_{m+1}(G_1, G_2). \qquad \square\end{aligned}$$

3. Relationship with edge ideal metrics

We shall show in this section that our independent set metrics are very related to the edge ideal metrics defined by Llabrés and Rosselló (2003). However, we start by showing that there are two main differences between the approaches leading to these metrics.

First, for a given n, we have presented a *finite* family of independent set metrics, as opposed to the infinite family of edge ideal metrics. Also, if we call the *independence*

number of a graph the size of its maximal independent set, and m is greater than both independence numbers of the two graphs, the independent set metrics are related with a simple scalar factor. Indeed, if m is greater than the independence numbers of G_1 and G_2, $\mathcal{I}(G_1)^{(m)} = \mathcal{I}(G_2)^{(m)} = \emptyset$, so $\mathcal{I}_m(G_1)\Delta\mathcal{I}_m(G_2) = \mathcal{I}_{m-1}(G_1)\Delta\mathcal{I}_{m-1}(G_2)$. Thus, $i'_m(G_1, G_2) = i'_{m-1}(G_1, G_2)$ and $\frac{a_{m-1}}{a_m} i_{m-1}(G_1, G_2) = i_m(G_1, G_2)$, for all graphs $G_1, G_2 \in \mathcal{C}_n$.

Second, Llabrés and Roselló suggest to calculate the Hilbert polynomial in order to evaluate their metrics. Doing so, any member of the metric family can be calculated with basically the same effort. However, to calculate the polynomial may be computationally demanding for big graphs. For instance, in Sloane's graph list (Sloane, 2003) the biggest graph for which CoCoA could calculate the Hilbert polynomial was one of 128 nodes and 1471 edges in 3 h. For bigger graphs it ran out of memory using 3 GB. Meanwhile, a simple brute force algorithm could count independent sets up to size 5 nodes on a graph of 256 nodes and 3839 edges in 3 min. Therefore, Hilbert polynomial calculation is not recommended for big graphs. We shall return to this issue in the next section.

In the rest of this section we prove that the evaluation of the Hilbert polynomial can be done by counting-independent sets and vice versa, and therefore, the metrics are very related.

First, let us review the basic definitions for the edge ideal metrics, with a more combinatorial and less algebraic flavour than in the original version.

Let $x_1, \ldots, x_n$ be variables. Let $\mathcal{M}(x_1, \ldots, x_n)$, or simply $\mathcal{M}(\underline{x})$, be the set of all monomials in the variables $x_1, \ldots, x_n$. We shall denote a monomial $x_1^{\alpha_1} \ldots x_n^{\alpha_n} \in \mathcal{M}(\underline{x})$ by $x^{(\alpha_1,\ldots,\alpha_n)}$ or simply by $x^{\underline{\alpha}}$ if we let $\underline{\alpha}$ stand for the n-tuple $(\alpha_1, \ldots, \alpha_n)$. The *total degree*, $|\underline{\alpha}|$, of a monomial $x^{(\alpha_1,\ldots,\alpha_n)}$ is $\sum_{i=1}^n \alpha_i$.

We say that x_i *divides* $x^{\underline{\alpha}}$, in symbols, $x_i | x^{\underline{\alpha}}$, when $\alpha_i \geq 1$, and analogously, that $x_i x_j$ *divides* $x^{\underline{\alpha}}$ when $\alpha_i,\ \alpha_j \geq 1$.

For every $m \geq 0$,

- let $\mathcal{M}(\underline{x})^{(m)}$ be the set of all monomials in $\mathcal{M}(\underline{x})$ of total degree m;
- let $\mathcal{M}(\underline{x})_m$ be the set of all monomials in $\mathcal{M}(\underline{x})$ of total degree $\leq m$.

Definition 3. For every graph,s $G = (V, E) \in \mathcal{C}_n$, the monomial set of G, $C(G)$, is the set of monomials in $\mathcal{M}(\underline{x})$ that are not divided by any pair of variables whose indexes form an edge in G:

$$C(G) = \{x^{\underline{\alpha}} | \text{ for all } \{i, j\} \in E,\ x_i x_j \not| x^{\underline{\alpha}}\}.$$

For every $m \geq 0$ and every graph G,

- let $C(G)_m = C(G) \cap \mathcal{M}(\underline{x})_m$ be the set of all monomials of total degree $\leq m$ that are in $C(G)$;
- let $H_G(m) = |C(G)_m|$ be the number of monomials of total degree $\leq m$ that are in $C(G)$.

We call this last mapping $H_G : \mathbb{N} \to \mathbb{N}$ the *Hilbert polynomial* of G and it is equal to the *Hilbert polynomial* $H_J : \mathbb{N} \to \mathbb{N}$, where J is the *edge ideal* associated to G (Llabrés and Roselló, 2003).

Now, we restate the relationship given by Llabrés and Roselló (2003) in their Lemma 12 between the square free monomials of a graph and the Hilbert polynomial and rewrite its proof to fit our notation. The following formula gives also a way to evaluate the Hilbert polynomial without calculating it.

Lemma 8. *Let $G = (V, E) \in \mathcal{C}_n$. For every $m \geq 0$,*

$$H_G(m) = \sum_{k=0}^{m} \binom{m}{k} \operatorname{ind}_G(k).$$

Proof. A monomial of the form $x_{i_1}^{\alpha_{i_1}}, \ldots, x_{i_k}^{\alpha_{i_k}}$, with $i_1, \ldots, i_k$ pairwise different and each $\alpha_{i_s} \geq 1$, belongs to $C(G)$ if and only if for every pair of variable indices i_s, i_t, $\{i_s, i_t\} \notin E$, which is equivalent to the fact that $\{i_1, \ldots, i_k\}$ is an independent set of G.

Therefore, each one of the $\operatorname{ind}_G(k)$ independent sets of size $k \geq 0$ adds as many monomials $x_{i_1}^{\alpha_{i_1}}, \ldots, x_{i_k}^{\alpha_{i_k}}$ to $C(G)_m$ as vectors $(\alpha_{i_1}, \ldots, \alpha_{i_k}) \in (\mathbb{N} - \{0\})^k$ such that $\sum_{t=1}^k \alpha_{i_t} \leq m$ there exist, and the number of the latter is $\binom{k+m-k}{k} = \binom{m}{k}$. Since all monomials in $C(G)_m$ added in this way are pairwise different by assumption, this proves that

$$|C(G)_m| = \sum_{k=0}^{m} \binom{m}{k} \operatorname{ind}_G(k),$$

as we claimed. □

To highlight the similarity implied by the previous lemma between the independent set metrics i'_m and the edge ideal metrics d'_m, remember that the metrics before normalisation satisfy that for every $G_1, G_2 \in \mathcal{C}_n$ and for every $m,\ 2 \leq m \leq n$,

$$i'_m(G_1, G_2) = I_{G_1}(m) + I_{G_2}(m) - 2I_{G_1 \cup G_2}(m),$$

$$I_G(m) = \sum_{k=0}^{m} \operatorname{ind}_G(k),$$

$$d'_{m+1}(G_1, G_2) = H_{G_1}(m) + H_{G_2}(m) - 2H_{G_1 \cup G_2}(m),$$

$$H_G(m) = \sum_{k=0}^{m} \binom{m}{k} \operatorname{ind}_G(k).$$

In words, I_G is calculated counting-independent sets, while H_G gives a different binomial weight to the independent set counters according to their relative size. In short, our

family is like theirs but without certain binomial terms, so that the absolute values calculated are smaller.

We could conclude from the previous observations that in order to evaluate Hilbert polynomials is better to first count-independent sets. However, this statement is not fair for two reasons. First, for some sparse graphs, Hilbert polynomials are much faster to calculate than to count-independent sets by brute force as we shall see in the next section. And, second, the following proposition says that improving the calculation of Hilbert polynomials can be used to improve independent set counting.

We will use that the formula in Lemma 8 can be written as

$$\mathrm{ind}_G(m) = H_G(m) - \sum_{k=0}^{m-1} \binom{m}{k} \mathrm{ind}_G(k)$$

and the following technical lemma on binomials.

Lemma 9. For all $m \geq 0$ and for any sequence $(b_i)_i$,

$$\sum_{k=0}^{m}\sum_{i=0}^{k}(-1)^{k-i}\binom{m+1}{k}\binom{k}{i} b_i = \sum_{i=0}^{m}(-1)^{m-i}\binom{m+1}{i} b_i.$$

Proof. Recall that by Newton's binomial formula,

$$\sum_{k=0}^{s}(-1)^k\binom{s}{k} = 0,$$

so

$$\sum_{k=0}^{m-i}(-1)^k\binom{m-i+1}{k} = (-1)^{m-i},$$

and

$$\sum_{k=i}^{m}(-1)^{k-i}\binom{m-i+1}{k-i} = (-1)^{m-i}.$$

We shall also use that

$$\binom{m+1}{k}\binom{k}{i} = \binom{m+1}{i}\binom{m-i+1}{k-i}.$$

Interchanging the sum order,

$$\sum_{k=0}^{m}\sum_{i=0}^{k}(-1)^{k-i}\binom{m+1}{k}\binom{k}{i} b_i = \sum_{i=0}^{m}\sum_{k=i}^{m}(-1)^{k-i}\binom{m+1}{k}\binom{k}{i} b_i$$

and using the equality of the binomials above,

$$= \sum_{i=0}^{m}\binom{m+1}{i} b_i \sum_{k=i}^{m}(-1)^{k-i}\binom{m-i+1}{k-i} = \sum_{i=0}^{m}\binom{m+1}{i} b_i(-1)^{m-i}. \qquad \square$$

Proposition 10. *The number of independent sets of cardinality m in G,* $\mathrm{ind}_G(m)$ *can be calculated using the values* $H_G(0), \ldots, H_G(m)$:

$$\mathrm{ind}_G(m) = \sum_{k=0}^{m}(-1)^{m-k}\binom{m}{k} H_G(k).$$

Proof. We prove it by induction. For $m = 0$, $\mathrm{ind}_G(0) = H_G(0)$.

Let us assume that the equality is true for m and we shall prove it for $m + 1$:

$$\begin{aligned} H_G(m+1) - \mathrm{ind}_G(m+1) &= \sum_{k=0}^{m}\binom{m+1}{k}\mathrm{ind}_G(k) \\ &= \sum_{k=0}^{m}\sum_{i=0}^{k}(-1)^{k-i}\binom{m+1}{k}\binom{k}{i} H_G(i) \\ &= \sum_{i=0}^{m}(-1)^{m-i}\binom{m+1}{i} H_G(i), \end{aligned}$$

using the previous lemma. From this, we get what we need:

$$\mathrm{ind}_G(m+1) = \sum_{i=0}^{m+1}(-1)^{m+1-i}\binom{m+1}{i} H_G(i). \qquad \square$$

4. Examples

In order to find all independent sets of size $\leq m$, we have used for simplicity a "brute force" (exhaustive search) algorithm that calculates them sorted by size and by the natural order of their vertices. First of all, we count the number of isolated nodes, remove them from the graph and use this value only at the end. An independent set of size j is used to find all independent sets of sizes $j + 1$ that have a new vertex greater than all vertices in the original independent set. The algorithm is polynomial as long as we consider m a fix number, and the algorithm is efficient as long as m is small ($\leq$6). This constrain is specially important when the graphs are sparse, as may be certain representation of biopolymers (see the first example below). For non-sparse graphs, as may be 3D neighbour structures (see the second example below), m could be a bit bigger.

We present some real and artificial graphs to show some relationship between the two families. The computational times are on a Pentium 4 (1.8 GHz, 250 MB of memory, 3 GB in disk).

4.1. Protein secondary structures

We represent part of the secondary structure of a protein as follows; see some schemes in Fig. 2. The peptides in a protein chain are represented by the vertices numbered in the chain order. An α helix is formed when each peptide in it has a hydrogen bond with the forth peptide ahead; thus, when there is an α helix in certain interval of the backbone, edges are included between node i and $i+4$ for as many nodes as the α helix length. A β sheet is formed when each peptide in one β strand has a hydrogen bond with a peptide in the next β strand. Therefore, edges $\{i, j\}$, $\{i-1, j+1\}, \ldots, \{i-l+1, j+l-1\}$ are added for anti-parallel β strands of length l, and analogously for parallel ones. All the information we need about these secondary structures is found in the protein description file of public databases like PDB (Berman et al., 2000).

We constructed graphs for 22 proteins with chains of length 100 in the database. We calculate four metrics for each family for all possible pairs. For space limitations, Table 1 shows only 20 of the 231 protein pairs; in this table, d_m refers to the edge ideal metric of type m.

We can see that the metrics are obviously related and we report that the edge ideal metrics are, on average, two times slower to calculate than the independent set metrics using a brute force algorithm for these very sparse graphs; the independent set distance values were calculated in 8 min. However, to calculate an additional member of the family based on Hilbert polynomials would take almost no additional time, while to calculate i_6 may double the time needed. In summary, for this type of sparse graphs, the Hilbert polynomial can be calculated faster than counting big size independent sets, although low order independent set metrics seem to be good enough.

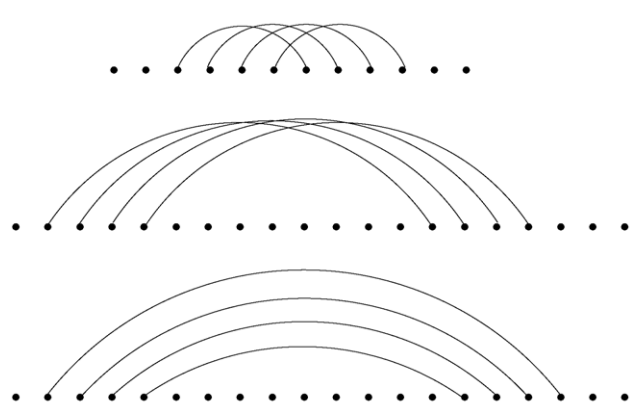

Fig. 2. Representation of an α helix, two parallel β strands and two anti-parallel β strands.

Table 1
Distances i_m vs.s distances d_{m+1}, $m = 2, \ldots, 5$

G_1	G_2	i_2	i_3	i_4	i_5	d_3	d_4	d_5	d_6
1c7m	1eod	41.0	40.5	39.7	38.6	41	40.5	39.7	38.7
1eod	1eoe	3.0	3.0	2.9	2.8	3	3	2.9	2.9
1eoe	1g2r	42.0	41.6	40.8	39.6	42	41.6	40.9	39.8
1g2r	1h0y	69.0	67.6	65.3	62.5	69	67.6	65.5	62.8
1h0y	1i6d	60.0	59.1	57.8	55.9	60	59.2	57.8	56.1
1i6d	1j0f	54.0	53.4	52.4	50.9	54	53.4	52.4	51.1
1j0f	1jbi	51.0	50	48.7	46.9	51	50	48.7	47.1
1jbi	1jnj	46.0	45.3	44.2	42.9	46	45.3	44.3	43.1
1jnj	1kq8	50.0	49.3	48.2	46.8	50	49.3	48.2	46.9
1kq8	1ksr	50.0	49.4	48.5	47.2	50	49.4	48.5	47.3
1ksr	1lds	37.0	36.5	35.7	34.6	37	36.5	35.7	34.7
1lds	1t1d	65.0	64	62.4	60.3	65	64	62.5	60.5
1t1d	1tnm	64.0	62.9	61.2	59	64	62.9	61.3	59.2
1tnm	1tnn	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1tnn	1uh6	61.0	59.9	58.3	56.3	61	60	58.4	56.5
1uh6	2ife	49.0	48.3	47.2	45.7	49	48.3	47.2	45.9
2ife	2lfb	60.0	59.1	57.7	55.8	60	59.1	57.8	56.1
2lfb	2psr	45.0	44.5	43.7	42.4	45	44.6	43.7	42.5
2psr	3crd	29.0	28.7	28.1	27.2	29	28.7	28.1	27.3
1c7m	3crd	36.0	35.7	35.2	34.5	36	35.7	35.3	34.6

One surprising experimental result is that if we plot i_m versus d_{m+1}, $m = 2, \ldots, 5$ (remember that d_{m+1} is defined in terms of monomials of degree up to m), we obtain an astonishing linear dependency (see Fig. 3). Besides, i_2 which is equal to d_3 because both correspond to the size of the symmetric difference of the set of edges, we were not expecting such a simple relation. Probably, the best explanation is that, for both metrics and in these examples, the value $\mathrm{ind}_G(m)$ is almost 100 times bigger than the sum of the other values for smaller m, so both metrics are using almost the same dominant value before normalisation. This is true for values of m that are low with respect to the independence number of the graphs.

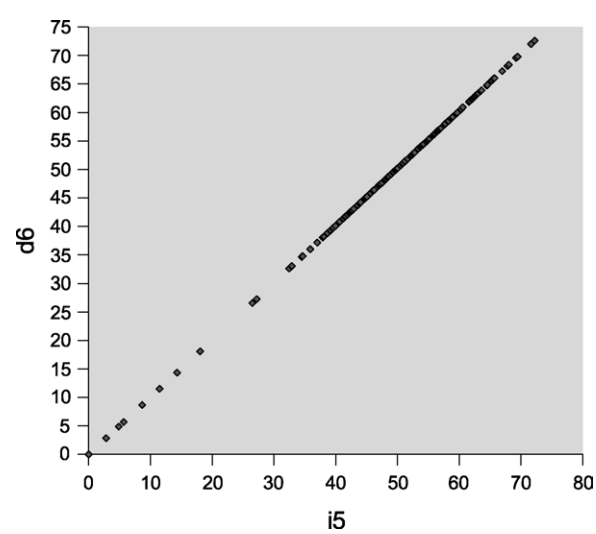

Fig. 3. The linear dependency between i_5 and d_6 for 231 protein pairs.

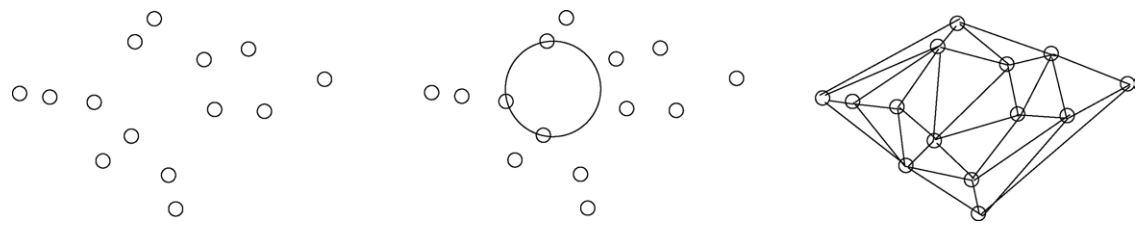

Fig. 4. On the left, a set of points. In the middle, three points that are neighbours because the shown circle does not contain any other point of the set. On the right, the Voronoi neighbours of this set.

4.2. Protein 3D structures

One way to represent the 3D structure of a protein is as a graph in which each node is an atom and nodes are related. if they are *Voronoi neighbours* in the following sense; two nodes located in three dimensions are Voronoi neighbours, if there exists a sphere that has both nodes on its surface and no other node in its interior. See a two-dimensional representation in Fig. 4. This notion is used in computer vision to represent "visibility" or "closeness" among points in a set, without using thresholds (Luo and Hancock, 2001). The representation is invariant to rotation and translation and is very stable to small relative motion of the nodes.

Coordinates of carbon atoms of a protein were extracted from the PDB files and the Voronoi neighbours of the atom set are calculated using qdelaunay, a program written for that purpose (Barber et al., 1996). For the protein 2A9L, 365 carbon atoms are extracted and 2600 arcs. The brute force algorithm could calculate the number of independent sets for up to size 5 in 1 h, while the corresponding Hilbert polynomial could not be calculated in 24 h of CPU.

Thus, the edge ideal metrics are not appropriate for comparing dense structures.

4.3. More examples

The graphs in Sloane's page cited in Section 3 show that the Hilbert polynomial is hard to calculate. It could be computed for the graph 1dc.128 in 3 h, time in which the brute force algorithm could count-independent sets until size 13, its independence number. Therefore, for not sparse graphs, independent set metrics of low-order could be calculated much faster than the Hilbert polynomial.

We also built a graph with 100 nodes and 100 edges that form one cycle. The Hilbert polynomial could not be calculated in 8 h of CPU time despite the fact that the graph is sparse. The brute force algorithm lasted 2 s. counting-independent sets of size up to 5.

5. Conclusion

We have shown a new family of metrics for graphs closely related to the edge ideal metrics but easier to describe and calculate. Thus, we argue that our metrics should be used instead for simplicity although our simple algorithm is slower for some sparse graphs. Should dense graphs that represent 3D protein structure need to be compared, we must count independent sets of small size since Hilbert polynomials cannot be calculated. Anyway, the new family still does not solve the problem of how to compare graphs with different number of vertices.

Acknowledgements

We thank F. Rosselló for having made comments on this manuscript that greatly improved it. This study was supported by grants BFM2003-00771 and INTAS IT 04-77-7178.

References

Barber, C., Dobkin, D., Huhdanpaa, H., 1996. The quickhull algorithm for convex hulls. ACM Trans. Math. Softw. 22 (4), 469–483, http://www.qhull.org.

Bayer, D., Stillman, M., 1992. Computation of Hilbert functions. J. Symbol. Comput. 14, 31–50.

Berman, A., et al. 2000. The protein data bank. Nucleic Acids Res. 28, 235–242, http://www.pdb.org.

Capani, A., Niesi, G., Robbiano, L., 2005. Cocoa, a system for doing computations in commutative algebra, available via anonymous ftp from http://www.cocoa.dima.unige.it.

Grayson, D.R., Stillman, M., 2005. Macaulay2, software system, available at http://www.math.uiuc.edu/Macaulay2.

Jou, M., Chang, G., 2000. The number of maximum independent sets in a graph. Taiwan. J. Math. 4 (4), 685–695.

Lesk, A.M., 2001. Introduction to Protein Architecture: The Structural Biology of Proteins. Oxfort University Press.

Llabrés, M., Rosselló, F., 2003. A new family of metrics for biopolymer contact structures. Comput. Biol. Chem. 28, 21–37.

Luo, B., Hancock, E., 2001. Structural matching using the EMalgorithm and singular value decomposicion. IEEE Trans. Pattern Anal. Mach. Intell. 23, 1120–1136.

Moulton, V., Zuker, M., Steel, M., Pointon, R., Penny, D., 2000. Metrics on RNA secondary structures. J. Comput. Biol. 7, 277–292.

Reidys, C., Stadler, P., 1996. Bio-molecular shapes and algebraic structures. Comput. Chem. 20, 85–94.

Schuster, P., Fontana, W., Stadler, P., Hofacker, I., 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proc. R. Soc. B 225, 279–284.

Sloane, N., 2003. Challenge problems: independent sets on graphs, http://www.research.att/.com/∼njas/doc/graphs.html.

Zuker, M., 1989. On finding all suboptimal foldings of an RNA molecule. Science 244, 48–52.