RUTCOR
RESEARCH
REPORT

USING LOGICAL ANALYSIS OF DATA (LAD) TO FIND PHYSIO-MECHANICAL DATA PATTERNS WHICH PREDICT CELLULAR OUTCOMES

Sascha Abramson[a] Gabriela Alexe[b]

Peter L. Hammer[c] Doyle Knight[d] Joachim Kohn[e]

RRR-40-2002 DECEMBER 2002

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/~rrr

[a] Department of Chemistry and Chemical Biology, Rutgers University, e-mail: sabrams@rci.rutgers.edu

[b] RUTCOR, Rutgers Center for Operations Research, Rutgers University, e-mail: alexe@rutcor.rutgers.edu

[c] RUTCOR, Rutgers Center for Operations Research, Rutgers University, e-mail: hammer@rutcor.rutgers.edu

[d] Mechanical and Aerospace Engineering, Rutgers University, e-mail: ddknight@rci.rutgers.edu

[e] Department of Chemistry and Chemical Biology, Rutgers University, e-mail: kohnpriv@rutchem.rutgers.edu

RUTCOR RESEARCH REPORT

RRR-40-2002 DECEMBER 2002

USING LOGICAL ANALYSIS OF DATA (LAD) TO FIND PHYSIO-MECHANICAL DATA PATTERNS WHICH PREDICT CELLULAR OUTCOMES

Sascha Abramson Gabriela Alexe

Peter L. Hammer Doyle Knight Joachim Kohn

Abstract. Although the exploration of predictive correlations between the physio-mechanical properties of biomaterials and cellular response is of critical importance to the design of clinically relevant medical devices, no mathematical models describing these relationships are known. This study represents an attempt to build a surrogate model of cellular growth in terms of physio-mechanical properties using a combinatorial library of highly structurally related degradable tyrosine-derived. The model was developed using a new knowledge extraction method for data analysis, based on combinatorics, optimization, and Boolean logic, called Logical Analysis of Data (LAD). The values of air-water contact angle, total flexibility index, and glass transition temperature of 62 polymers were used to find patterns that led to specific cellular outcomes (metabolic activity). In spite of the limited size of the training set, the model was validated by correct prediction of the metabolic activity of cells grown on 5 of 6 previously untested materials. Most critically, the model was able to find patterns of physio-mechanical properties that led to good cellular response. With these patterns a single set of criteria was established that described all the "superior" materials in the study. These patterns are critical in the design of new biomaterials as they make possible the screening of materials in the computer rather than on the benchtop.

Keywords: Logical Analysis of Data, LAD, Polyarylate, Combinatorialpublicly available datasets. In the last part of the paper, we present the results of a series of computational experiments which show the high degree of robustness of spanned patterns.

Acknowledgements: The partial support provided by ONR grant N00014-92-J-1375 is gratefully acknowledged.

Introduction

While many predictive correlations between chemical structure and the physio-mechanical properties of polymers have been established, based on sound, fundamental understanding of the behavior of polymeric materials, a comparable body of knowledge relating to the behavior of living cells is not yet available. In particular, appropriate mathematical models to describe cellular behavior to different substrata do not yet exist. Currently, this information about a particular material can only be gathered by conducting extensive cell growth experiments on that material. Both costly and time-consuming, this approach is impractical when large numbers of test materials have to be evaluated, nor does it give any global understanding of the relationship of physio-mechanical properties of biomaterials to specific cellular outcomes.

In order to address these challenges, the research presented here represents an attempt to develop a mathematical model capable of predicting cellular responses to polymeric biomaterials as a function of their physio-mechanical properties. Additionally, we identified “patterns" of properties of these polymers, which led to specific biological responses. The use of these patterns can substantially reduce the number of materials that have to be laboratory tested, both by identifying lead materials, and eliminating others. Additionally, the patterns substantially increase our understanding of biomaterials as a whole.

Recent developments in combinatorial design of biomaterials have led to a library of tyrosine-derived polyarylates (Kohn 1994). This library contains 112 strictly alternating copolymers of a diacid and a diphenol component (Figure 1). By systematically varying the structure, certain physio-mechanical properties, such as air water contact angle (θ), total flexibility index (TFI), and glass transition temperature (Tg), could also be systematically changed. Initial studies found these three properties for all 112 polymers in addition to normalized metabolic activity (NMA) of fibroblasts grown on 62 to the polymers (Brocchini, James et al. 1998).

A mathematical model was developed using a knowledge extraction method for data analysis, based on combinatorics, optimization, and Boolean logic, called Logical Analysis of Data (LAD). Initially created for the classification of binary data (Hammer 1986; Crama, Hammer et al. 1988), LAD was later extended (Boros, Hammer et al. 1997) to datasets having binary (positive and negative) observations, depending on numerical variables. The central concepts used by LAD are those of positive and negative patterns, which are special sets (or conjunctions) of conditions imposed on the values of the variables. A conjunction is called a positive pattern if, on the one hand, a sufficiently high proportion of the positive observations in the dataset satisfy it (are covered by it), and, on the other hand, each negative observation violates at least one of the conditions in the definition of that conjunction.

The fundamental assumption of LAD, validated on dozens of datasets, is that the knowledge of the patterns can be inferred to unseen cases. The collection of all patterns (or the “pandect”) hidden in a dataset can be enormously large because of its inherent redundancies (Alexe and Hammer 2001; Hammer, Kogan et al. 2001; Alexe, Alexe et al. 2002; Alexe and Hammer 2002). In order to reduce the redundancies in the pandect, and to increase the explanatory power of the classification models, LAD selects a minimal subset of positive patterns and one of negative patterns. A LAD model consists of two sets of patterns, constituting respectively, a positive and a negative theory.

Models present a convenient way for classifying new observations. If a new observation "triggers" a pattern in the positive theory and none in the negative theory, it will be classified as being positive. Conversely, a new observation triggering a pattern in the negative theory and none in the positive theory will be classified as negative.

While the prediction of yet unseen observations is at the core of many modeling schemes, LAD's main advantage is to go beyond the aims of prediction (or classification). One of the most important pieces of additional information provided by LAD is the explanatory element given by the patterns. The pattern-based LAD models are completely reproducible, have a high explanatory power, and, due to a built-in "space discretization" mechanism, are almost insensitive to missing or to noisy data.

LAD has been successfully applied for the analysis of datasets in medicine (identification of cardiac patients at high/low risk (Alexe, Blackstone et al. 2002; Lauer, Alexe et al. 2002), ovarian (Alexe, Alexe et al. 2002) and breast (Alexe, Alexe et al. 2002) cancer detection), economy (Hammer, Hammer et al. 1999), finance (Alexe, Hammer et al. 2002), oil exploration (Boros, Hammer et al. 2000), and seismology (Boros, Hammer et al. 2000). Computational studies (Alexe, Alexe et al. 2002) showed that the accuracy of the LAD models compared favorably with the accuracy of other machine learning and statistical models.

In the present study, a LAD model was constructed using a training dataset of 62 polymers. In spite of the small size of the training set, LAD identified a number of patterns, leading to a model which was able to predict biological outcome with a high accuracy. The simplicity of the LAD patterns provide them with clear explanatory power. The predictive power of the model was validated by randomly choosing six polymers from the 50 not included in the training set, and growing cells on them. The experimental data closely matched the predicted values, thus validating this method. Most critically the LAD model found a pattern of positive and negative data that described all superior materials in the study.

Materials and Methods

Library design and synthesis

A previously published data set of 112 tyrosine-derived polyarylates was selected for examination by LAD. Briefly, in the novel combinatorial approach to polymer design strictly alternating A-B copolymers were synthesized from one diphenol monomer (A) with a series of reactive pendent chains (Figure 2) and a second dicarboxylic acid monomer (B) which allows for systematic changes in the backbone (Figure 3). By utilizing 14 diphenols and 8 dicarboxylic acids in all possible combinations 112 individual polymer compositions resulted (Brocchini, James et al. 1997; Brocchini, James et al. 1998).

The air-water contact angle (θ), total flexibility index (TFI), and glass transition temperature (Tg) was measured for each of the 112 polyarylates. Additionally, cellular response was measured by an in vitro cell metabolic assay and normalized to metabolic activity (NMA) of cells grown on tissue culture polystyrene (TCPS) (Brocchini, James et al. 1997).

Data analysis

The general LAD methodology described by Boros, et al (Boros, Hammer et al. 2000) has been adapted to the specific features of this dataset (e.g., numerical rather than binary outcome, sharply limited number of observations). Data was represented in the 12 dimensional space associated to the 3 numerical variables (TFI, θ, and Tg), and to the 2 categorical variables corresponding to the pendent chains type (1 if DTE or DTM, 0 otherwise), and to the backbone type: adipate, digylocolate, dioxaoctanedioate, glutarate, methyl adipate, sebacate, suberate, and succinate, respectively; for example, poly(DTD succinate) was represented as the 12-vector (16 96 40 0 0 0 0 0 0 0 0 1).

The analysis consisted of several stages. After the identification of two outliers in the dataset, and eliminating them from further consideration, thresholds maximizing the combinatorial contrast between the groups of observations having NMA values below or above these thresholds were determine. Using these thresholds, the categories of "strong", "medium", and "weak" observations were defined. Further, the numerical variables were discretized, by identifying a minimal set of cutpoints. Next, various collections of maximal positive and negative patterns were produced, and filtered in order to identify minimal positive and negative theories. Leave-one-out cross-validation procedures were applied then to models consisting of a positive and a negative theory, in order to establish the optimal control parameters (degree, prevalence) for theory and model construction. Finally, a LAD model was selected and validated on "a test set "of yet unseen observations, as described below.

Validation

Six of the remaining 50 polymers in the library were randomly chosen to validate the LAD model. The NMA of fibroblasts on these polymers was measured as previously (Brocchini, James et al. 1997). Briefly, the polymers were spin coated onto 15 mm glass. Once dry the coverslips were loaded polymer side up into 24-well polystyrene plates. Rat lung fibroblasts (RFL-6) cells were grown on the coverslips for 7 days and metabolic activity was measured by a commercially available kit (CellTiter96®, Promega, Madison, WI). Metabolic activity was normalized to tissue culture polystyrene. Experimental values were compared to the values predicted by the LAD.

Results and Discussion

Data consistency

We have found that the internal consistency of the training data is very high, and only a small number of observations were discordant. The few inconsistencies we have identified are due perhaps either to measurement errors, or to the existence of additional parameters that were not included in the data. The main reason for which we suspect that a few observations may contain mistakes is that if we represent them (in normalized form) in the 12 dimensional space of the numerical and categorical variables, it can be seen that their closest neighborhoods contain points having sharply different NMA values. For instance, close neighborhoods of poly(DTB dioxaoctanedioate) and poly(DTM sebacate), which have NMA values of 83 and 89, include respectively, poly(DTB adipate) and poly(DTB methyl adipate), having sharply differing NMA values (32 and 35, respectively). For similar reasons, we suspect that poly(DTE methyl adipate), poly(DTiP sebacate), and poly(DTB suberate) may also contain mistakes. A consequence of the inconsistency of the five observations mentioned above with the rest of the data set is the fact that they are not covered by most of the powerful patterns (although some of them may be covered by very weak ones). To increase the quality of the LAD model, we eliminated from the training set two of the most inconsistent outliers: poly(DTB dioxaoctanedioate), and poly(DTM sebacate).

Class identification

Polymers with high NMA values were distinguished from those having low NMA values by constructing a model for identifying strong and weak observations, i.e. those which have NMA values above or below a certain threshold. Since polymers with an NMA value above the range 70-80 can be viewed as having a very good cellular response and those with an NMA values below the range 50-60 can be viewed as having a poor cellular response, we experimented with several models corresponding to several threshold values t and t' in these ranges. The positive theory P for t = 77 (Table 1) and the negative theory N for t' = 55 (Table 2) provided the best combinatorial separation of the strong class from the weak one and therefore, the chosen model was based on these thresholds. The polymers with intermediary NMA values (above 55 and below 77) may display properties specific for both weak and strong classes, and consequently, they are considered as a "buffer" of medium value. Figure 4 illustrates the NMA ranges we considered in this study.

The LAD model

In order to separate the strong and the weak classes, LAD produced a model consisting of a positive theory P which distinguishes the strong observations (those with NMA ≥ 77), and a negative theory N which distinguishes the weak observations (those with NMA ≤ 55) (Figure 4). The LAD model consists of 11 patterns which are presented in Tables 1 and 2. The columns of Table 1 correspond to the 5 positive patterns $P_1,\ldots, P_5$ in the model, while the columns of Table 2 correspond to the 6 negative patterns $N_1,\ldots., N_6$ in the model; these columns describe the pattern defining constraints. The 12 lines in the tables correspond to the variables (features) recorded in the dataset. The last line in tables represents the prevalence of the patterns, i.e. the proportion of the 16 strong and 21 weak observations satisfying each pattern. It can be seen that the degree of each of the patterns in the model (i.e. the number of variables in its defining conditions) is at most 3.

Analysis of variables

Some important observations were made by examing the patterns LAD. All 5 variables in the dataset (TFI, θ, Tg, pendent chains, and backbone) play an important role in determining the value of NMA. This conclusion based to the facts specified values of TFI, θ and Tg appear in each of the patterns and the removal of pendent chains from the LAD model, was seen to lead to an 11% drop of accuracy, while the elimination of backbone from the same model decreased its accuracy by 20%. The higher values of TFI and θ can be seen to be indicative of lower NMA values. On the contrary, higher values of Tg can be seen to be indicative of higher NMA values. Pendent chains of type DTE or DTM are almost entirely excluded by the negative patterns, i.e. they are indicative of higher values of NMA. The pattern describing the superior polymers reinforces all the above conclusions. Indeed, these polymers have very low TFI and contact angle values, high glass transition, and have the pendent chains either DTE or DTM.

Superior polymers

By definition, an observation that satisfies at least one of the positive patterns in the model has a high NMA value. We found a group of positive observations, to be called superior observations, which satisfy all of positive patterns. The NMA values of the 6 observations in this group (poly(DTE glutarate), poly(DTE diglycolate), poly(DTM diglycolate), poly(DTM glutarate), poly(DTE succinate), poly(DTM succinate)) are of 78, 82, 88, 95, 97, and 115, respectively, with an average of 92.5.

It is also remarkable to know that there exists a special pattern Π that distinguishes completely the superior observations from all the other ones:

$$(\Pi): \quad \begin{cases} \text{TFI} \le 7 \\ \text{è} \le 71 \\ \text{Tg} \ge 67 \\ \text{Pendent chains are either DTE, or DTM} \\ \text{Backbone is neither adipate, dioxaoctanedioate, methyl adipate, sebacate, nor suberate} \end{cases}$$

More exactly, the superior observations satisfy each of the 5 constraints that define Π, and every other observation violates at least one of them.

The pattern Π could be used in addition to the LAD model, as a powerful set of conditions indicative of polymers with excellent cellular response, suggesting that only those polymers should be considered for further evaluation in in vivo experiments.

Classification and validation

The LAD model can be used for classification. It can be seen that each of the 21 observations having NMA values below 55 triggers some of the negative patterns in the model, and none of the positive ones. Similarly, 16 out of the 18 observations having NMA values above 77 trigger some of the positive patterns, and none of the negative ones; the remaining 2 uncovered observations (poly(DTB dioxaoctanedioate) and poly(DTM sebacate)) are in fact the outliers detected and disregarded by LAD. None of the medium observations trigger any pattern characteristic for the strong or for the weak class.

Thus the model allows for the correct classification of all the weak and all the strong consistent observations in the training set, leaving uncovered only the two outliers disregarded by LAD. Moreover, taking into account the fact that the dataset contains only 62 observations, far less than in previous applications of LAD, the existence of positive patterns with high prevalence (ranging from 56% to 81%) indicates also the reliability of LAD-based classifications for yet unseen polymers.

When validating the model on the test set of 6 new observations, we found that the predictions were correct in 5 cases (83%). It should be added that only one of the 6 test polymers, poly(HTE diglycolate), which has the NMA value 109, is classified as superior by Π. The only error in the predictions concerned poly(HTE adipate) having NMA value 67.1, which was incorrectly classified as strong, although the model recognized that it is not superior. Table 3 presents the validation results on the test set.

Conclusions

The LAD technique achieved a remarkably accurate set of predictions of cell metabolic activity on a series of tyrosine-derived polyarylates, in spite of the small number of observations in the dataset. The model was trained using a dataset consisting of 62 of the 112 polyarylates for which a set of characterization data (materials properties and cell growth) was available. In the process of training the LAD model, 11 individual patterns were identified. Each of these patterns related the ranges of input data to either a very high (strong) or a very low (weak) value of cell growth on the test polymers. These patterns were used to create a model, which fully characterized all cell growth data for the weak and strong polymers in the training set, leaving only 2 outliers unclassified. The reliability of the LAD model was validated on 6 randomly chosen polymers from the remaining 50 that were not used in the training. The specific LAD predictions grouped the expected cell growth outcomes (more precisely the expected NMA values) for the so-far untested polymers into discrete relative categories of NMA behavior ("strong", “medium”, "weak"), and recognized the polymers belonging to the “superior” category. In validation tests, LAD correctly predicted the metabolic outcome for 5 of the 6 polymers. Moreover, the LAD model discovered a class of superior polymers, characterizing it by the simultaneous fulfillment of 5 conditions. These patterns can be used to screen new materials for lead compounds for further investigation or eliminate poor candidates from further research. Ultimately, these techniques will be able to save time and resources in the search for new biomaterials for clinical applications.

References

Alexe, G., Alexe, S., Axelrod, D., Boros, E., Hammer, P.L. (2002). “Combinatorial analysis of breast cancer data from image cytometry and gene expression microarrays.” Manuscript.

Alexe, G., Alexe, S., Hammer, P.L., Kogan, A. (2002). "Comprehensive vs. comprehensible classifiers in Logical Analysis of Data." Piscataway, New Jersey, RUTCOR, Rutgers Universtiy Research Report RRR 9-2002, Ann Oper Res (in print).

Alexe, G., Alexe, S., Hammer, P.L., Liotta, L., Petricoin, E., Reiss, M. (2002). "Ovarian cancer detection by logical analysis of proteimic data." Manuscript.

Alexe, G., Hammer, P. L. (2002). "Spanned patterns in the Logical Analysis of Data." Piscataway, NJ, RUTCOR, Rutgers Universtity Research Report RRR 15-2002 Ann Oper Res (in print).

Alexe, S., Blackstone, E., Hammer P.L., Ishwaran H., Lauer M.S., Snader C.E.P. (2002). "Coronary risk prediction by Logical Analysis of Data." Ann Oper Res **115** (in print).

Alexe, S., Hammer, P. L. (2001). "Accelerated algorithm for pattern detection in Logical Analysis of Data." Piscataway, New Jersey, RUTCOR, Rutgers University Research Report RRR 21-2001**,** Ann Oper Res (in print).

Alexe, S., Hammer, P. L., Kogan, A., Lejeune, M. (2002). "Decoding Standard & Poor's country risk-rating." Piscataway, New Jersey, RUTCOR, Rutgers University Research Report RRR 2002.

Boros, E., Hammer, P. L., Ibaraki T., Kogan A. (1997). "A logical analysis of numerical data." Math Program **79**: 163-190.

Boros, E., Hammer, P. L., Ibaraki T., Kogan A., Mayoraz, E., Muchnick, I. (2000). "An implementation of Logical Analysis of Data." IEEE Trans Knowledge Data Eng **12**: 292-306.

Brocchini, S., James, K., et al. (1997). "A combinatorial approach for polymer design." J. Amer. Chem. Soc. **119**(19): 4553-4554.

Brocchini, S., James, K., et al. (1998). "Structure-property correlations in a combinatorial library of degradable biomaterials." J. Biomed. Mater. Res. **42**: 66-75.

Crama, Y., Hammer, P. L., Ibaraki T. (1988). "Cause-effect relationships and partially defined Boolean functions." Ann Oper Res **16**: 299-326.

Hammer, A., Hammer, P. L., Muchnick, I. (1999). "Logical analysis of Chinese productivity patterns." Ann Oper Res **87**: 165-176.

Hammer, P. L. (1986). "Partially defined Boolean functions and cause-effect relationships." International Conference on Multi-Attribute Decision Making Via OR-Based Expert Sytems, University of Passau, Passau Germany.

Hammer, P. L., Kogan, A., Simeone, B., Szedmak S. (2001). "Pareto-optimal patterns in Logical Analysis of Data." Piscataway, New Jersery, RUTCOR, Rutgers University Research Report RRR 7-2001, Discr Appl Math (in print)

Kohn, J. (1994). "Tyrosine-based polyarylates: Polymers designed for the systematic study of structure-property correlations." 20th Annual Meeting of the Society for Biomaterials, Boston MA, Society for Biomaterials.

Lauer, M. S., Alexe S., Blackstone E., Ishwaran H., Hammer P.L. (2002). “Use of the Logical Data Analysis method for assessing long-term mortality risk after exercise electrocardiography.” Circulation **106**: 685-690.

Figure Legends

Figure 1. General chemical structure of the library of tyrosine-derived polyarylates. These materials are alternation A-B co-polymers consisting of a diacid and a diphenol component. The polymers can be systematically varied at Y and R (for additional details see Figures 2 and 3).

Figure 2. The chemical structure of the diphenols used to create structure variation at the polymer pendent chain

Figure 3. The chemical structure of the diacids used to create structural variations in the polymer backbone

Figure 4. Ranges for the expected cell growth NMA values. Polymers with very low (below 55) NMA values constitute the weak class, while those with very high (above 77) NMA values constitute the strong class. Polymers having the NMA between 55 and 77 may display simultaneously properties specific for strong and weak class, and constitute a buffer of medium value.

Figures

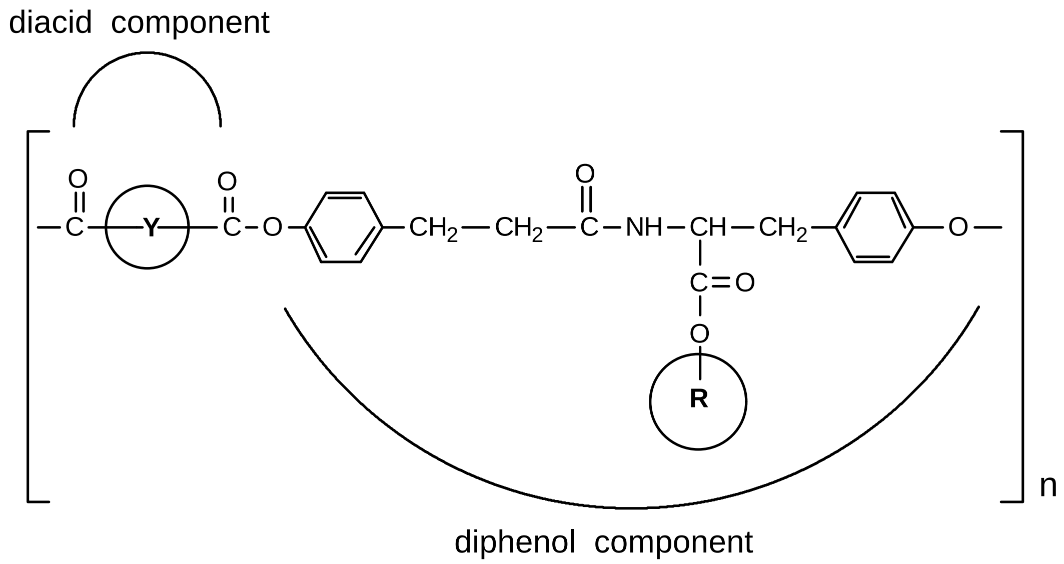

Figure 1

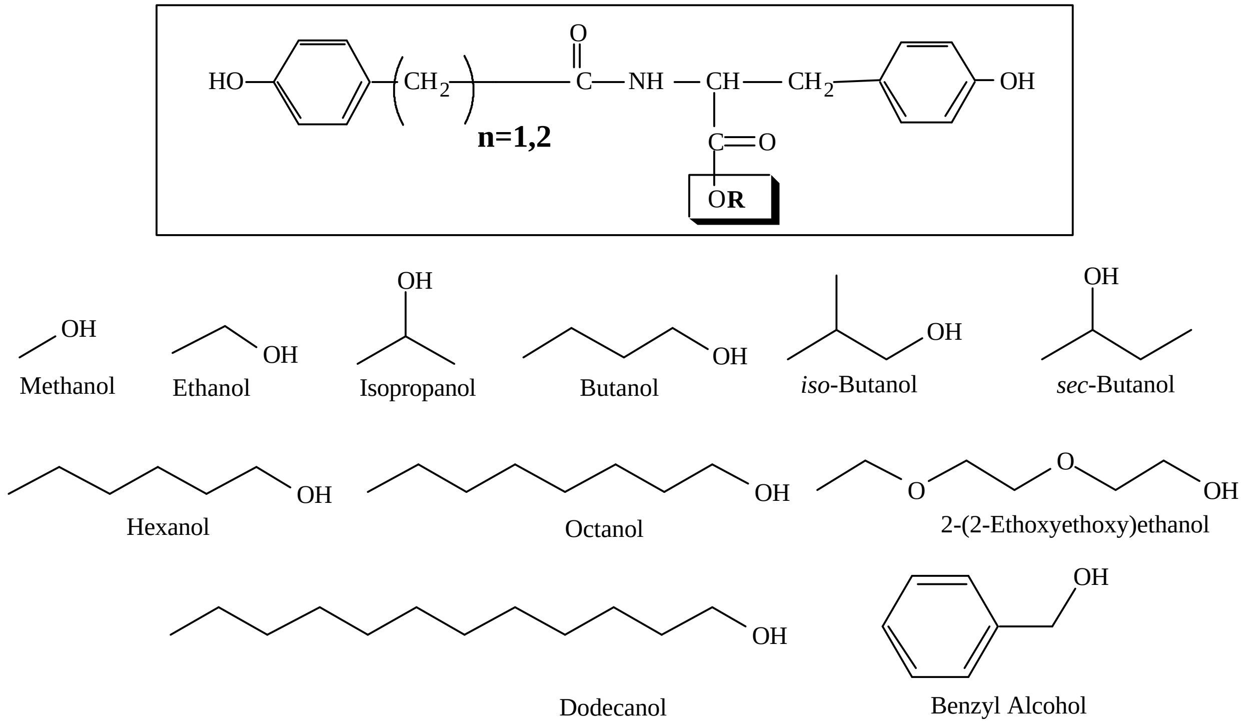

Figure 2

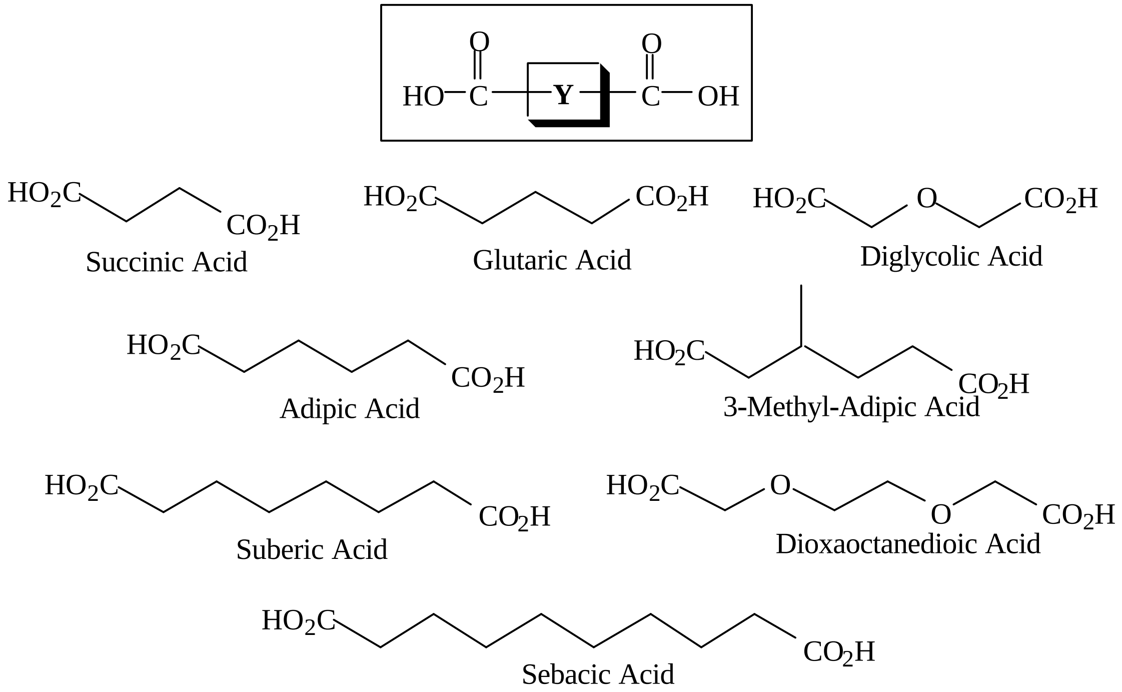

Figure 3

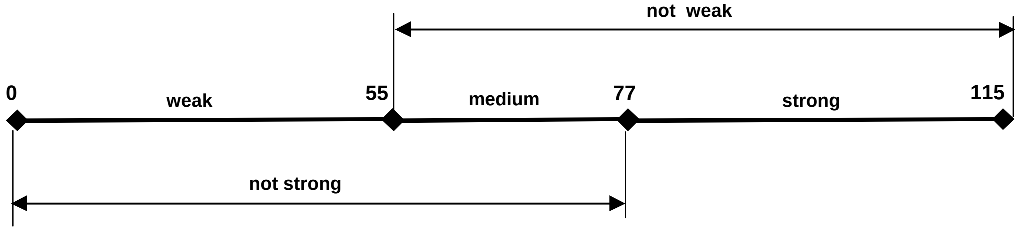

Figure 4

Tables

Table 1. Positive theory P: pattern description.

		Patterns				
		P1	**P2**	**P3**	**P4**	**P5**
Physio-mechanical Descriptors	TFI	≤10	≤10	≤11	≤8	
	θ	≤72	≤73	≤73		≤74
	Tg			> 63	> 67	
Pendent Chains	DTE or DTM					Yes
Backbone	Adipate		No			No
	Digylocolate					
	Dioxaoctanedioate					
	Glutarate					
	Methyl Adipate					
	Sebacate					
	Suberate					
	Succinate					
Prevalence (% of the 16 strong observations satisfying the pattern)		81	75	75	69	56

Table 2. Negative theory N: pattern description.

		Patterns					
		N1	**N2**	**N3**	**N4**	**N5**	**N6**
Physio-mechanical Descriptors	TFI		> 9				
	θ	> 86			> 77	> 74	
	Tg	≤ 31					> 40, ≤ 64
Pendent Chains	DTE or DTM		No	No	No	No	
Backbone	Adipate		Yes				
	Digylocolate						
	Dioxaoctanedioate	No					
	Glutarate				Yes		
	Methyl Adipate			Yes			Yes
	Sebacate						
	Suberate						
	Succinate					Yes	
Prevalence (% of the 21 weak observations satisfying the pattern)		43	24	19	14	14	14

Table 3. Validation results of LAD model. `

Polymer	Variables			NMA	Patterns Triggered		LAD Prediction		Correctness of LAD Predictio
	TFI	θ	Tg		Positive	Negative	Class	Superior	
Poly(DTiB adipate)	10	77	56	41.4		N2	Weak	No	Yes
Poly(HTE adipate)	7	71	65	67.1	P1, P3		Strong	No	No
Poly(DTiB diglycolate)	9	74	72	62.5			Medium	No	Yes
Poly(HTE diglycolate)	6	66	66	101.5	P1, P2, P3, P5, Π		Strong	Yes	Yes
Poly(HTH glutarate)	10	83	42	53		N4	Weak	No	Yes
Poly(HTH methyl adipate)	12	83	38	63.7			Medium	No	Yes