

Contents

I	Prologue	1
1	Justification, Objectives and Contributions	3
1.1	MultiScaleS and SynBioFactory projects	3
1.2	Objectives of this thesis	6
1.3	Contributions	10
1.3.1	Articles in peer-reviewed journals	10
1.3.2	Conference contributions	11
1.3.3	Software	13
1.3.4	Awards	13
2	On chemometrics	15
2.1	Introduction	16
2.2	Notation	16
2.3	Exploratory data analysis	18
2.3.1	Principal component analysis (PCA)	18
2.3.2	Missing-data methods for exploratory data analysis (MEDA)	20
2.3.3	Maximum likelihood principal component analysis (MLPCA)	21
2.3.4	Multivariate curve resolution (MCR)	22
2.4	Regression models	23
2.4.1	Principal component regression (PCR)	23
2.4.2	Partial least squares regression (PLS)	23
2.4.3	Joint-Y PLS (JYPLS)	25
2.5	Missing data	26
2.5.1	PCA model building and model exploitation with missing data	28
2.5.2	PLS model building and model exploitation with missing data	29
2.6	<i>N</i> -way data	30
2.6.1	<i>N</i> -way Partial least squares regression (NPLS)	31
3	On systems biology	33
3.1	Introduction	34
3.1.1	Systems biology: a paradigm shift	34
3.1.2	Origins	36
3.1.3	Aim and goals	36

3.2	Genomics and transcriptomics	37
3.3	Proteomics	38
3.4	Metabolomics and fluxomics	39
3.4.1	First principle models	40
3.4.2	Stoichiometric modelling	41
3.4.3	Network-based pathway analysis: elementary modes (EMs)	43
3.4.4	Possibilistic consistency analysis	44
3.5	For every omic science: Network inference	47
3.5.1	Mutual information distance and entropy reduction (MIDER) method	47
4	Material	51
4.1	Hardware	51
4.2	Software	51
4.3	Biological organisms	52
4.4	Datasets	53
II	Modelling biological organisms	55
5	Metabolic flux understanding	57
5.1	Introduction	58
5.2	<i>Pichia pastoris</i> metabolic model	60
5.2.1	Metabolic network reconstruction	60
5.2.2	Experimental data set	60
5.3	Grey modelling	62
5.3.1	Possibilistic consistency analysis	62
5.3.2	MC sampling	65
5.4	Multivariate modelling	66
5.4.1	PCA with MEDA	66
5.4.2	MCR-ALS	69
5.5	Discussion	76
5.6	Conclusions	78
5.7	Appendix. Metabolic model.	78
6	Projection to elementary modes	83
6.1	Introduction	84
6.2	Principal elementary modes analysis	84
6.2.1	Data preprocessing	87
6.2.2	Algorithm	88
6.3	Case studies	88
6.3.1	<i>E. coli</i> simulated study	88
6.3.2	<i>E. coli</i> real data	91
6.3.3	<i>Pichia pastoris</i> real data	96

6.4	Discussion	100
6.5	Conclusion	100
6.6	Appendix A. PEMs.	101
6.7	Appendix B. Metabolic models.	101
7	Dynamic elementary mode modelling	117
7.1	Introduction	118
7.2	Metabolic models of <i>Saccharomyces cerevisiae</i>	119
	7.2.1 Metabolic networks	119
	7.2.2 Concentration data	119
7.3	Dynamic elementary mode modelling	122
	7.3.1 Dynamic elementary mode analysis (dynEMA)	122
	7.3.2 Dynamic elementary mode regression discriminant analysis (dynEMR-DA)	124
7.4	Triple cross-validation procedure (3CV)	126
7.5	Results	128
	7.5.1 Simulated flux data	128
	7.5.2 Actual flux data	129
7.6	Discussion	132
7.7	Appendix. Metabolic models.	133
8	Fusing different omics data sources	137
8.1	Introduction	138
8.2	Potyvirus and its proteins	140
8.3	Protein-Protein Interaction Network (PPIN) reconstruction	141
8.4	Mutations and fitness	143
8.5	Mathematical modelling	146
8.6	Statistical modelling	147
8.7	Functional modules	150
8.8	Discussion and conclusions	153
9	Multivariate image analysis for fruit discrimination	155
9.1	Introduction	156
9.2	Experiment	158
9.3	Methodology	159
	9.3.1 Data preprocessing	159
	9.3.2 Feature extraction	160
	9.3.3 Discriminant models, validation procedure and wavelength selection	161
9.4	Results	166
9.5	Conclusions	168
9.6	Appendix. Preprocessings.	169

III	Missing data	171
10	PCA model building with missing data	173
10.1	Introduction	174
10.2	Methodology	174
10.3	Data sets	182
10.4	Comparative study	182
10.5	Results	184
10.5.1	Olive Oil data set	184
10.5.2	Diesel data set	185
10.5.3	Simulated data set	185
10.5.4	Big data set	186
10.6	Discussion and conclusions	187
10.7	Appendix. Methods equivalences.	188
11	Network inference with missing data and outliers	191
11.1	Introduction	192
11.2	Methods	194
11.2.1	Missing data methods	194
11.2.2	Outlier detection and correction	194
11.2.3	Case studies	196
11.3	Results	197
11.3.1	Missing data: comparative study	197
11.3.2	Outlier detection and correction: a simulated study	200
11.3.3	Remark on computation times	201
11.4	Discussion	201
12	Missing data imputation toolbox	205
12.1	Introduction	206
12.2	Software specifications and requirements	206
12.3	Data sets	207
12.4	Operating procedure	207
12.5	Concluding remarks	212
12.6	Appendix A. Excel files.	213
12.7	Appendix B. Using MATLAB command window.	213
13	Framework for MLPCA missing data imputation	215
13.1	Introduction	216
13.2	Maximum likelihood regression-based methods	217
13.3	Data sets	220
13.4	Comparative study	220
13.5	Results	221
13.5.1	FTIR microspectroscopy	221
13.5.2	<i>P. pastoris</i> cultures on heterogeneous culture media	221

13.5.3	Simulated data set	223
13.5.4	Additional data sets	224
13.6	Conclusions	225
13.7	Appendix A. Regression-based imputation step in MLPCA.	225
13.8	Appendix B. Additional figures.	226
14	Calibration transfer between near infrared instruments	229
14.1	Introduction	230
14.2	Materials	231
14.3	Methods	231
14.3.1	Piecewise direct standardisation (PDS)	231
14.3.2	MLPCA and TSR	232
14.3.3	JYPLS	232
14.4	Modelling procedure	233
14.5	Results	236
14.5.1	Gasoline dataset	236
14.5.2	Corn dataset	240
14.6	Discussion	241
14.7	Conclusions	244
15	PLS model building with missing data	247
15.1	Introduction	248
15.2	Adaptations of TSR for PLS-MB	249
15.2.1	From PCA model building (TSR-1)	250
15.2.2	From PLS model exploitation (TSR-2)	250
15.3	Data sets	253
15.4	Comparative study	255
15.5	Results	256
15.5.1	Hald data	256
15.5.2	<i>P. pastoris</i> data	256
15.5.3	NIR data	258
15.5.4	Simulated data	258
15.6	Discussion and conclusion	261
IV	Epilogue	265
16	Conclusions	267
16.1	Meeting the objectives	267
16.2	Relevance	271
16.3	Future lines	272
	Bibliography	275
	Abbreviations and acronyms	307